Examination of the Parameter Estimate Bias When Violating the Orthogonality Assumption of the Bifactor Model

By

Chunmei Zheng

Submitted to the Department of Psychology and Research in Education and the
Faculty of the Graduate School of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

Neal Kingston, Ph.D., Chairperson

William P. Skorupski, Ph.D.

Bruce B. Frey, Ph.D.

Carol M. Woods, Ph.D.

John Rury, Ph.D.

Date defended: ______________________

The Dissertation Committee for Chunmei Zheng certifies
that this is the approved version of the following dissertation :

Examination of the Parameter Estimate Bias When Violating the Orthogonality Assumption of the Bifactor Model

Neal Kingston, Ph.D., Chairperson

Date approved: ______________________

Abstract

Educational and psychological constructs are normally measured by multifaceted dimensions. The measured construct is defined and measured by a set of related subdomains. A bifactor model can accurately describe such data with both the measured construct and the related subdomains. However, a limitation of the bifactor model is the orthogonality assumption that proposes that there are no correlations between the measured construct and its associated subdomains and among the associated subdomains (i.e., orthogonal). This assumption requires that all items perfectly measure the specified constructs so that no correlations exist among all the constructs. In other words, test developers need to write items perfectly, measuring the primary and one subdomain factor only. However, test items are inherently flawed in practice and can rarely be written perfectly. To force correlated factors to be orthogonal can result in a loss of information and can lead to distorted and untrustworthy parameter estimates in bifactor solutions. Precision of parameter estimates is an important issue in any assessment because parameter estimates are considered to be decisive criteria for finalizing item performance and respondents' ability level. The purpose of this study, therefore, was to investigate the parameter estimate bias of different levels of orthogonality violation among factors of the bifactor model. Since the orthogonality violation cannot be controlled and true parameters are unknown in real data, an extensive series of simulation studies were generated to evaluate a proposed bifactor model with various orthogonality violations among the subdomains. Results indicated that levels of orthogonality violation had no significant influence on intercept and theta parameter estimates but did have a significant influence

on discrimination parameter estimates. Higher levels of orthogonality violation could result in severely distorted discrimination parameter estimates. Orthogonality violations between two subdomains could only distort parameter estimates of the involved subdomains and the primary construct but not the other subdomains. Among all of the theta parameter estimates, the estimates of the primary dimension were most trustworthy. Item length had no significant influence on either item or person parameter estimates.

Acknowledgements

I would like to first thank my family for their strong support. I will never forget your encouragement, your financial support, your help and care for my son. Without your support, I would not have been able to accomplish my goal and obtain my Ph.D. I am proud of being a member of this family. Second, I would especially extend my gratitude to Dr. Neal Kingston, my adviser and dissertation chair. Your insightful guidance during my doctoral study has made me well prepared and competent in this field. Your thoughtful comments, corrections, and unlimited encouragement truly helped me to complete my work. Thanks also to Dr. William P. Skorupski for your generous knowledge input about statistics and test theory, and Dr. Bruce B. Frey, Dr. Carol M. Woods, and Dr. John Rury for your valuable comments. I am very honored to have you serving as my committee members. Third, I want to specifically thank Dr. Paul Johnson. I have learned so much from you regarding data generation and program writing. Thanks so much for spending your valuable time helping me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

Educational and psychological constructs, also named as factors, traits, or abilities, are normally measured by multifaceted dimensions (Immekus & Imbrie, 2008; Seo, 2011). The measured construct is defined and measured by a set of related subdomains. For example, math ability can be measured by number, algebra, geometry, and data factors as indicated by the assessment framework from the Kansas Department of Education, and the self-determination ability can be measured by autonomous, self-realization, and empowerment factors as indicated by the Arc Self-Determination Scale from the National Longitudinal Transitional Study-2. Data collected from these assessment instruments are multidimensional; they have one primary dimension, the measured construct, and several related subdomains for specifying the primary factor. For instance, the primary dimension in a math test is math ability, while the subdomains are number, algebra, geometry, and data factors. Similarly, the primary dimension in the self-determination scale is the self-determination ability, and the subdomains are autonomous, self-realization, and empowerment factors. To accurately describe or make inferences about such data, a model needs to represent the structure with both the primary and the related specific factors.

A variety of item response models can be used to evaluate such data. These are unidimensional, multidimensional, and bifactor item response models (e.g., Chen, West, & Sousa, 2006; Ebesutani et al., 2011; Patrick, Hicks, Nichol, & Krueger, 2007; S. P. Reise, Morizot, & Hays, 2007). Unidimensional item response theory models assume that all items of an assessment instrument measure only one single common construct, which explains the intercorrelations among all items. Unidimensional item response models are easy to implement. The analysis can be conducted by a number of statistical software programs such as BiLOG-MG, Testfact, WinBUGS, OpenBUGS, Mplus, and IRTpro. However, it quickly became apparent that the unidimensionality assumption was often violated because assessments often consist of multiple dimensions.

When multidimensional data are estimated with a unidimensional model, only the primary dimension is taken into consideration, as was the case in the early works of item response theory. The unidimensional item response models can only provide estimates about the latent trait level of an examinee on the primary factor the instrument was designed to measure. The multifaceted subdomains are ignored. Ignoring the subdomains means ignoring the internal item clusters which are likely to be more correlated among themselves than items between clusters, so the local dependence assumption that items are uncorrelated after accounting for the latent construct of item response theory is violated. DeMars (2006) observed that "violation of local independence can lead to overestimates of reliability or information, underestimates of the standard error of the ability level, and it also can lead to mis-estimation of item parameters" (p. 147). Unidimensional item response models are reasonable for tests with only one strong construct.

It is possible to analyze multidimensional data with a multidimensional model. Multidimensional item response models assume that multiple dimensions exist among test items. That is, to correctly answer a test item, participants need knowledge of multiple constructs. Multidimensional data can be estimated by two types of multidimensional item response models: a compensatory or a noncompensatory model. A compensatory multidimensional item response model means that lower ability in one dimension can be offset by higher ability in another dimension. That is, a compensatory model means that probability of an incorrect response can be compensated by higher values of another trait level. Conversely, a noncompensatory multidimensional item response model means that lower ability in one dimension cannot be helped by higher ability in another dimension (Bolt & Lall, 2003; Reckase, 2009). High probability of success for a compensatory model requires high proficiency in some dimensions, while high probability of success for a noncompensatory model requires high proficiency in all dimensions. Dichotomously-scored data such as multiple-choice items with either right or wrong answers can be estimated by both compensatory and noncompensatory multidimensional item response models. However polytomously-scored data, such as Likert scale items, can only be estimated with compensatory item response theory models. Noncompensatory multidimensional item response models have not been proposed so far (Reckase, 2009).

Multidimensional item response models are becoming more and more popular in the literature because they take the complexity of psychological and educational constructs into consideration. The models can truthfully model the multidimensionality of multi-domain data. In addition, scores on subdomains can be derived and then used for diagnostic purposes. For instance, lower scores of a subdomain construct such as algebra indicate that students may not have mastered the knowledge of the concept. Educators can work on that by providing extra assistance or more exercises. However, multidimensional item response models can only estimate parameters of subdomains not of the primary construct. S. P. Reise et al. (2007) said that when items tend to have small loadings on the primary factor and larger loadings on group factors, multidimensional item response models should be used. Therefore, multidimensional item response models are appropriate only when items have small loadings on a primary construct but large loadings on subdomain constructs.

When multidimensional data are estimated by a bifactor item response model, both primary dimension and subdomains are taken into account. In this regard, bifactor item response models are particularly useful for capturing the multidimensional data with both a primary dimension and subdomains, the sort of data commonly encountered in educational and psychological measurement. Bifactor item response models assume that all items measure a primary construct explaining intercorrelations among all the items. In addition, the models also specify two or more subdomains. These subdomains explain variance not accounted for by the primary factor. Unlike hierarchical models in which subdomains are modeled as first order and the primary factor as the second order, all of the factors in bifactor item response theory models are in the first order. Researchers can obtain information related to item performance and latent trait abilities for all dimensions, and statistical analysis can also be conducted directly on all dimensions. The primary ability estimates are useful for important decisions such as accountability, while the subdomains complement the overall ability estimate by providing finer diagnosis of examinees' strengths and weaknesses (De la Torre, Song, & Hong, 2011). A theoretical bifactor pattern of eight hypothesized items with one primary factor and four subdomains can be described as:

$$a = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ a_{31} & 0 & a_{33} & 0 & 0 \\ a_{41} & 0 & a_{43} & 0 & 0 \\ a_{51} & 0 & 0 & a_{54} & 0 \\ a_{61} & 0 & 0 & a_{64} & 0 \\ a_{71} & 0 & 0 & 0 & a_{75} \\ a_{81} & 0 & 0 & 0 & a_{85} \end{bmatrix}$$

Where a_{jk} represents the loadings of item j (j = 1, 2, 3...J) on latent factor k (k = (1, 2, 3...K). The figure indicates that all of the eight items load on the primary factor (a_{j1}), as well as on one of the subdomains (a_{j2}, a_{j3}, a_{j4}, and a_{j5}). For example, the first two items measure not only the primary dimension but also the first subdomain, and the second two items measure the primary dimensions and the second subdomain. Likewise, the third and fourth item pairs measure the primary and corresponding third and fourth factors.

The bifactor item response models can be applied to both dichotomously- and polytomously-scored data. For dichotomously-scored data, it can be estimated by a two-parameter logistic or a three-parameter logistic bifactor model. A two-parameter logistic bifactor model estimates discrimination (a) and difficulty (b) parameters, while a three-parameter logistic bifactor model estimates discrimination, difficulty, and guessing parameters (c). Item discrimination, which is analogous to factor loading, refers to how well an item discriminates respondents with high standing from respondents with low standing on a latent factor. Item difficulty, also known as item threshold or item location, can be regarded as the likelihood with which a respondent will select a correct or a more positive response. The guessing parameter refers to the probability of a correct answer when simply guessing. For polytomously-scored data, only the two-parameter logistic bifactor model is appropriate because item response differs only in intensity. For example, “agree” and “strongly agree” in a Likert scale differ only in intensity. Since there is no right or wrong answer for polytomously-scored data, there is no pure guessing in the actual answers. The guessing parameter for polytomously-scored data normally will not be estimated.

In recent years, bifactor item response models have been increasingly applied to empirical data from both achievement tests and multi-domain psychological instruments. Compared to other models such as unidimensional item response models, multidimensional item response models, and factor-analytic structures, bifactor models indicate a superior relative model fit (e.g., Brouwer, Meijer, Weekers, & Baneke, 2008; Brown, Finney, & France, 2011; Golay & Lecerf, 2011; Martel, Von Eye, & Nigg, 2010; Reininghaus, McCabe, Burns, Croudace, & Priebe, 2011; Simms, Grös, Watson, & O’Hara, 2007). Reise, Moore, and Haviland (2010) discussed the advantages of bifactor item response models. First, bifactor models provide parameter estimates on a primary construct while controlling for the possible distortion of the multidimensional nature of data. Therefore, bifactor models count for the effect of item clusters that unidimensional item response models ignore. Second, the models provide informative statistics such as discrimination indices and ability scores for the subdomains. The information can be used for diagnostic purposes to improve item and examinee performance. Third, the models can assist in the study of parameter estimate distortion when estimating multidimensional data with unidimensional item response models.

1.2 Statement of the Problem

Even though the bifactor item response model has many advantages, it is not without limitations. One major restriction of the bifactor item response model is the orthogonality assumption (Chen et al., 2006; Holzinger & Swineford, 1937; S. P. Reise, Horan, & Blanchard, 2011; Simms et al., 2007). The orthogonality assumption proposes that there are no correlations between the primary factor and its associated subdomains and among the associated subdomains themselves. The bifactor factors are orthogonal or independent of each other. Consequently, the current bifactor item response model is also called a restricted (Reise et al., 2010), traditional (Martel et al., 2010), or canonical version of a bifactor model (Chen et al., 2006). Bifactor models have two crucial features: first, data are measuring a primary and a number of subdomain constructs; and second, the constructs are orthogonal or independent of each other. The first feature of this model requires multidimensional data, which can be met frequently because many researchers hypothesize a primary construct that encompasses several related subdomains in psychological and educational assessments. But the orthogonality assumption that requires all items to perfectly measure the constructs of an instrument is difficult to meet in terms of practical item development.

According to Reise et al. (2010), "the restricted bifactor model demands not only that data be multidimensional, but also that the multidimensional data be well structured (i.e., each item measures a general trait and one and only one subtrait)" (p. 557). In other words, test developers need to write items perfectly, measuring only the primary factor and only one subdomain. The problem is whether it is possible to write such perfect items in reality. Again taking math ability as an example, if the math construct being tested is measured by four subdomains such as numbers, algebra, geometry, and data, the orthogonality assumption requires test developers to write neatly packaged units that function independently of one another. It is difficult to imagine how one can correctly answer the algebra items without number knowledge, or correctly answer data items without number knowledge. There will be cross loadings among these items. One could conclude that the latent factors can theoretically be assessed separately, but in reality, there will be some overlap. Forcing correlated factors to be orthogonal can result in a loss of information and can lead to distorted and untrustworthy item parameter estimates in the restricted bifactor solution.

To summarize, the orthogonality assumption can produce interpretable results, but this is a flawed argument. It can rarely be possible both practically and theoretically. In light of the aforementioned points, it is unsuitable to impose the orthogonality assumption on multidimensional data with a primary dimension and a number of subdomains. To impose an orthogonality model with nonorthogonal data may cause problems such as item and person parameter estimates bias.

1.3 Purpose and Structure of the Study

Little has been investigated of the parameter estimate bias when violating the orthogonality assumption. Precision of parameter estimates is an important issue in any assessment, because parameter estimates are considered to be decisive criteria for finalizing item performance and respondents' ability level. The purpose of this study, therefore, was to investigate the parameter estimate bias of different levels of orthogonality violation among subdomains of a bifactor model. Data were simulated to be dichotomous with either a right or wrong answer, and only a two-parameter bifactor model was implemented. The selection of a two-parameter bifactor model with dichotomously-scored data was made for the following two reasons: two-parameter biactor models are easy to understand and implement with software, and dichotomously-scored data are most common in educational assessments. Fixed factors of this present study were a sample size of 5,000 participants, five dimensions with one primary dimension and four related subdomains, and an equal number of items in all subdomains. Manipulated factors included the number of items and levels of orthogonality violation. An empirical study with real data was applied with the bifactor item response analysis to explore the possible parameter estimate bias. Specifically, the following research questions were addressed:

(1) How are the item and person parameter estimates different from the true parameters under various levels of orthogonality violation? The bias between the estimated and the true parameters

were computed, and then descriptive statistics such as the mean, the standard deviation, and the quartiles of the bias scores were reported to examine the recovery of parameter estimates.

(2) Do the number of items and models with orthogonality violations among different subdomains result in a significant estimate bias of item and person parameters? The number of items was 40, 60, and 80, and the models refers to orthogonality violation between two subdomains as Model 1 and orthogonality violation among all subdomains as Model 2. The mean bias and the Root Mean Square Error (RMSE) were reported across different numbers of items and models. Analysis of variance (ANOVA) was then performed to examine whether number of items and models were significant predictors of parameter estimate bias.

(3) Are all item and person parameters equally off from the true parameters, or will orthogonality violations between two subdomains affect only the involved dimensions or other dimensions as well? The mean bias and the mean RMSE were also reported across all dimensions and specific models. ANOVA was also performed to examine if dimensions and specific models are significant predictors of parameter estimate bias.

The independent variables in the bifactor simulation study were models with violations between two subdomains and among all subdomains, levels of orthogonality violation, and number of items. The dependent variables were item and person parameter estimate bias. It was hypothesized that parameter estimates of the orthogonal and trivial violation would lead to insignificant differences between the estimated and the true parameters. However, the more severe the orthogonal assumption violation, the more distorted the parameter estimates. The structure of this dissertation is as follows. Chapter 2 briefly describes the history, the current bifactor application, and the limitations of a bifactor item response model. Chapter 3 presents the research design of the simulation study that assesses parameter estimate bias. The results of the simulation study and a real data example are presented in Chapter 4. Chapter 5 summarizes and discusses the simulation results in the context of educational and psychological assessment. In addition, limitations of the study are described and future directions for research are considered.

1.4 Study Terminology

There are some abbreviations and terms in the present study. The following definitions are provided to ensure uniformity and understanding of these terms throughout the study. The researcher developed all definitions.

- Int: Intercept parameters of a bifactor model
- P: The general or primary factor of a bifactor model
- S1: The first subdomain of a bifactor model
- S2: The second subdomain of a bifactor model
- S3: The third subdomain of a bifactor model
- S4: The fourth subdomain of a bifactor model
- Model 1: A bifactor model when orthogonality violations exist between the third and the fourth subdomains
- Model 2: A bifactor model when orthogonality violations exist among all subdomains
- Model 11: The level of orthogonality violations between the third and the fourth subdomains is 0.1
- Model 12: The level of orthogonality violations between the third and the fourth subdomains is 0.2
- Model 13: The level of orthogonality violations between the third and the fourth subdomains is 0.4
- Model 14: The level of orthogonality violations between the third and the fourth subdomains is 0.6
- Model 21: The level of orthogonality violations among all subdomains is less than 0.1

- Model 22: The level of orthogonality violations among all subdomains is greater than or equal to 0.1 but less than 0.3

- Model 23: The level of orthogonality violations among all subdomains is greater than or equal to 0.3 but less than 0.5

- Model 24: The level of orthogonality violations among all subdomains is greater than 0.5

- Models: Model 1 and Model 2

- Specific Models of Model 1: Model 11, Model 12, Model 13, and Model 14

- Specific Models of Model 2: Model 21, Model 22, Model 23, and Model 24

1.5 Summary

Different models can be used to model multidimensional data with a primary factor and a number of subdomain factors. They are unidimensional, multidimensional, and bifactor item response models. The bifactor model has been increasingly applied in psychological and educational assessments because it can accurately capture the data structure with a primary and several subdomains. The overall ability estimate is of major interest to most educators and practitioners because it often serves as an important criterion with which to make final decisions about student achievement, while subdomain estimates of the ability level serve to complement the primary ability level estimates by providing information about respondents' strengths and weaknesses. The bifactor model can best represent this multidimensional data and provide more information than other models.

However, there is a major restriction with the bifactor model: the correlations between the primary factor and its subdomains and among subdomains are constrained to be orthogonal. The model assumes there are no correlations among all the latent factors, but this constraint is unrealistic. Constructs drawn from substantive complex measures are often correlated with each other. To impose orthogonality on correlated factors can lead to untrustworthy item and person parameter estimates. The purpose of this study, therefore, was to examine the parameter estimate bias of different orthogonality violations. Results reported parameter estimate bias and effects of the number of items, models, dimensions, and specific models on parameter estimate bias.

Chapter 2

Literature Review

Bifactor item response models are a special type of multidimensional item response models, which are an extension of unidimensional item response models. The underlying theory of unidimensional models is that all items of an assessment measure a single factor. However, it has long been recognized that few psychological assessments are strictly unidimensional. Many measures are constructed in such a way that items are clustered to reflect different aspects of a trait (e.g., Chen et al., 2006; Ip, 2010). When item response data are not unidimensional, forcing data into a unidimensional model will lead to distorted and, thus, untrustworthy parameter estimates. Consequently, many research studies have examined the robustness of parameter estimates under different levels of unidimensional violation. Another way to solve this problem is to estimate multidimensional data with a multidimensional model. Therefore, a unidimensional item response theory model is extended to a multidimensional model to describe different dimensions.

Bifactor models are a special case of multidimensional models because bifactor models contain more than one latent construct. Multidimensional models can be categorized as simple structure or complex structure. Multidimensional models with the simple structure happen when items load on the dimensions they are designed to measure, whereas the multidimensional model with a complex structure happens when some items load on more than one dimension. The bifactor model, thus, is a complex-structure multidimensional model. Bifactor models extract a factor beyond the different dimensions as the primary factor and make the different dimensions uncorrelated (S. P. Reise et al., 2007).

This chapter synthesizes and discuses relevant research studies of a bifactor model. It first introduces the history of a bifactor model and then discusses the applications of the bifactor model in research studies. The history and the current practice are presented to indicate the development of the bifactor model. The model fit, specification, and sample size of the bifactor model are discussed respectively at the end.

2.1 History of the Bifactor Model

The bifactor model was initially developed by Holzinger and Swineford (1937) as an extension of Spearman's two-factor pattern. Their study briefly introduced the bifactor method and also illustrated how this method may be modified for the analysis of variables of greater complexity. In Spearman's conceptualization of cognitive abilities, all variables are related to a general factor and yet each contains a specific factor. A similar structure also appeared in the interbattery factor analysis of Tucker (1958), while a confirmatory factor analysis model was considered by (Jöreskog, 1969). Bock and Aitkin (1981) introduced marginal maximum likelihood estimation of parameter estimates. Kingston and McKinley (1988, April) developed a new model, confirmatory multidimensional item response theory model, to assess the dimensionality of mental test data. Gibbons and Hedeker (1992) developed full-information item bifactor analysis for binary item responses. The article described how parameters of the item bifactor model for binary responses can be estimated by maximum marginal likelihood using a variation of the EM algorithm described by Bock and Aitkin (1981). Gibbons et al. (2007) recently extended the bifactor model to the case of polytomous items (e.g., multi-category rating scales).

2.2 Bifactor Model Application

So far, bifactor models have been applied to address dimensionality and item performance issues for educational and psychological assessments. Researchers have solved the issues by comparing different competing models. The competing models include unidimensional, multidimensional with multiple factors, hierarchical models, or second-order models. By examination of model fit such as log-likelihood, chi-square, and p value, researchers have made conclusions about the dimensions of an assessment instrument. Within an item response theory framework, item performance has also been examined (Martel et al., 2010; Norris, 2010; Swineford, 1941; Thomas, 2012; Thomas & Locke, 2010; Weekers, 2009). Some researchers further explored the advantages of the bifactor model over other models (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Chen et al., 2006; S. P. Reise, Ventura, et al., 2011).

Bifactor item response models were also applied to examine differential item functioning (DIF) items with a Bayesian approach. Fukuhara (2009) developed a new DIF detection model by extending the bifactor model with a Bayesian Markov Chain Monte Carlo estimation method. Results of the simulation study indicated that a bifactor item response DIF model can estimate DIF and non-DIF items accurately, and can also classify non-DIF items more accurately compared with other item response DIF models. The bifactor DIF model can properly estimate the DIF magnitude and had low DIF detection error rates.

Some researchers touched upon estimation algorithm of the bifactor model. Yang, Song, and Xu (2002) described a robust estimator for correlated observations based on bifactor equivalent weight. The robustness and efficiency of the new estimator was demonstrated with simulation data. Gibbons et al. (2007) extended full-information item factor analysis from a binary response model to a graded response model. Marginal maximum likelihood estimation of item parameters was utilized for graded response bifactor analysis. Cai (2010a) described a two-tier full-information item factor analysis model with a development of an EM algorithm for full-information marginal maximum likelihood estimation. He indicated that multidimensional item response theory models, bifactor models, and testlet response models are special types of the two-tier full-information framework. Cai (2010b) also introduced the Metropolis-Hastings Robbins-Monro algorithm for exploratory and confirmatory item factor analysis. Results reported the advantages of this algorithm over existing methods and its accuracy and efficiency in parameter estimates.

Besides confirmatory bifactor item response theory models, research studies have also addressed exploratory bifactor analysis with orthogonal and oblique cases. The exploratory bifactor model, under the orthogonal case, studies item and person parameters by assuming the latent variables are not correlated. Under the oblique case, the assumption of orthogonality is relaxed. Item and person parameters are examined by letting the latent variables correlate. Schmid and Leiman (1957) and Jennrich and Bentler (2011) developed techniques for exploratory bifactor model under orthogonality assumption. Schmid-Leiman's orthogonalization technique has been widely used in exploring dimensions of different psychological instruments (Reise et al., 2010; S. P. Reise, Horan, & Blanchard, 2011; S. P. Reise, Ventura, et al., 2011). Jennrich and Bentler's technique is new, and the relations between their method and Schmid-Leiman's orthogonalization were discussed in the article. In the same year, Jennrich and Bentler (2011) developed technique for oblique case.

Bifactor models have been applied to testlet-based assessment as well (DeMars, 2006; Rijmen, 2009, 2010). DeMars compared four models: the bifactor model, the testlet-effects model, the polytomous model, and the independent-items model. Differences among models were reported with simulated data. Rijmen (2009) described three multidimensional IRT models: bifactor model, testlet model, and second-order model. This showed that second-order models are equivalent to testlet models and both second-order and testlet models are special cases of the bifactor model.

Absolute model-data fit of bifactor models have been discussed in the literature. Li and Rupp (2011) described the absolute model fit statistic, s-χ^2. Most studies applied relative model fit when comparing models as well as information-based relative fit indices, such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC), to select a better-fitting model among a set of nonnested models. The author examined the Type I error rates and power of the model fit s-χ^2 statistic for the full-information bifactor model under various simulation conditions. The author also studied how the AIC, BIC, and corrected Akaike's information criterion (CAIC) indices for relative model fit perform under different conditions of model misspecification involving unidimensional item response models, multidimensional item response models, and full-information bifactor models.

Because bifactor models are mostly limited to a single-group analysis and dichotomously-scored data, Cai, Yang, and Hansen (2011) proposed a multi-group bi factor model in which the latent mean and variance of each group can be estimated and compared across groups, and items can be an arbitrary mixing of dichotomous, ordinal, and nominal type. An effective full-information maximum marginal likelihood estimator was derived, and it achieved substantial computational savings by extending the bi factor dimension reduction method.

2.3 Bifactor Model Fit

The structure of bi factor models consists of a primary domain and several sub domains. It assumes that correlations among all of the dimensions, the primary and the several sub domains, are orthogonal. Researchers view this bi factor structure with suspicion. Limitations of the bi factor structure have been discussed by researchers such as Chen et al. (2012, 2006) and Reise et al. (2010); S. P. Reise et al. (2007). The limitations of the structure are summarized below. First, the orthogonality assumption is problematic in reality. It requires test developers to write perfect items that measure only the designed construct, however, it is possible that the items contain concepts of multiple domains. Therefore, it is unrealistic to assume the orthogonality assumption from the test development perspective. In addition, it is also hard to interpret the orthogonality correlations among the latent factors. In other words, the bi factor structure actually suggests that there are correlations among the latent factors because the primary factor is specified by several related sub domains. The actual implementation of the bi factor models states that factors are independent of each other instead. Reise et al. (2010) said that "some researchers are skeptical that the model itself makes any sense" (p. 557).

2.4 Bifactor Model Specifications

Model specification is another problem. Reise et al. (2010) stated that in order to make the model identifiable and obtain trustworthy parameters, the bi factor model should contain at least three sub domain factors, the numbers of items should be equal for each sub domain factor, and there should be at least three items for each sub domain factor. Questions then arise when there are two subdomains but the data are multidimensional. In this case, unidimensional models are not appropriate, and bifactor models are not identifiable. Also, what about when there are three subdomains, but there are only two items in the third subdomain? Will parameter estimates be accurate? And will the dimensionality be recovered? Furthermore, Reise et al. (2010) concluded that the model is “too clumsy for routine use in structural modeling” (p. 557) because bifactor models need to estimate parameters for substantially more paths than other models. Some researchers argue that the solution represents an overdetermination of the data.

2.5 Bifactor Model Sample Size

Use of the bifactor model necessitates a large sample size due to large parameter estimates. Chen et al. (2012) found that “As any structural equation models (SEMs), bifactor models require sufficiently large sample size” (p. 245). The minimum sample size may depend on several factors, but the sample size is definitely a factor to consider when estimating bifactor models.

In conclusion, bifactor models are a natural progression of item response theory development. Unidimensional item response models are developed to examinees’ trait level by considering the relationship between item and person parameters. The model indicates that the probability of success is a function of item parameters (e.g., discrimination and difficulty parameters) and a person parameter (i.e., the latent trait level). But researchers or practitioners quickly realize that psychological assessments are seldom unidimensional. Often, there is built-in multidimensionality in these assessments. Multidimensional models have been developed in response to the complexity of the actual assessments. The bifactor model is a type of multidimensional model. It considers both the primary dimension (like unidimensional models) and the subdomains (like multidimensional models). Bifactor models are appropriate when there is a strong primary as well as strong subdomain constructs.

The bi factor item response theory model has gained more attention and has been increasingly applied to empirical data from both achievement tests and psychological instruments due to its superior relative model fit. It also has more advantages over other models such as uni dimensional, multidimensional, and second-order models in representing dimension issues and obtaining informative statistics such as item discrimination or difficulty induces. However, it is not without limitations. It assumes orthogonality among factors, it requires a large sample size, and it needs at least three sub domain factors and balanced numbers of items in each subdomain in order to make the model identifiable. The orthogonality assumption is a flawed statement because it can seldom be met in reality. Forcing orthogonality can result in distorted and untrustworthy parameter estimates. Little has been examined about the parameter estimate bias when violating this assumption. The purpose of the present research, therefore, is to examine the parameter estimate bias when violating the orthogonality assumption. In addition, factors such as the number of subdomains, balanced number of items, and sample size are all taken into consideration because they can influence the implementation of the bifactor model.

Chapter 3

Method

The purpose of the present research was to study the parameter estimate bias when the orthogonality assumption between and among subdomains of a bifactor model was violated. Because the orthogonality violation can not be controlled and true parameters are unknown in real data, an extensive series of simulation studies were generated to evaluate a proposed two-parameter bifactor item response theory model with various levels of orthogonality violations between and among the subdomains. The proposed bifactor model had one primary dimension and four related subdomains, which were measured by an equal numbers of items. There were 5,000 participants, and items were dichotomously-scored with either right or wrong answers. The fixed factors were the five dimensions, an equal number of items in each subdomain, and a total of 5,000 participants. Manipulated factors included numbers of items and levels of orthogonality between and among subdomains. This chapter describes the simulation design, data generation, evaluation criteria, summary of procedures, and an empirical study with real data.

3.1 Simulation Design

Data were generated based on three bifactor models (see Figures 3.1, 3.2, and 3.3). The Validity Model is ideal because it satisfies the orthogonality assumption that no correlations exist among all latent factors. It was used to test the accuracy of the bifactor program for simulation studies written in *R*. Model 1 indicates an orthogonality violation between two subdomains due to cross loadings. The third and fourth subdomains were selected to indicate levels of the orthogonality violation. It is of particular interest to discover whether the violation between two subdomains influences the parameter estimates of all dimensions, or only the involved dimension. Model 2 reveals an orthogonality violation across all subdomains. Results can indicate the parameter estimate bias under different levels of orthogonality violation.

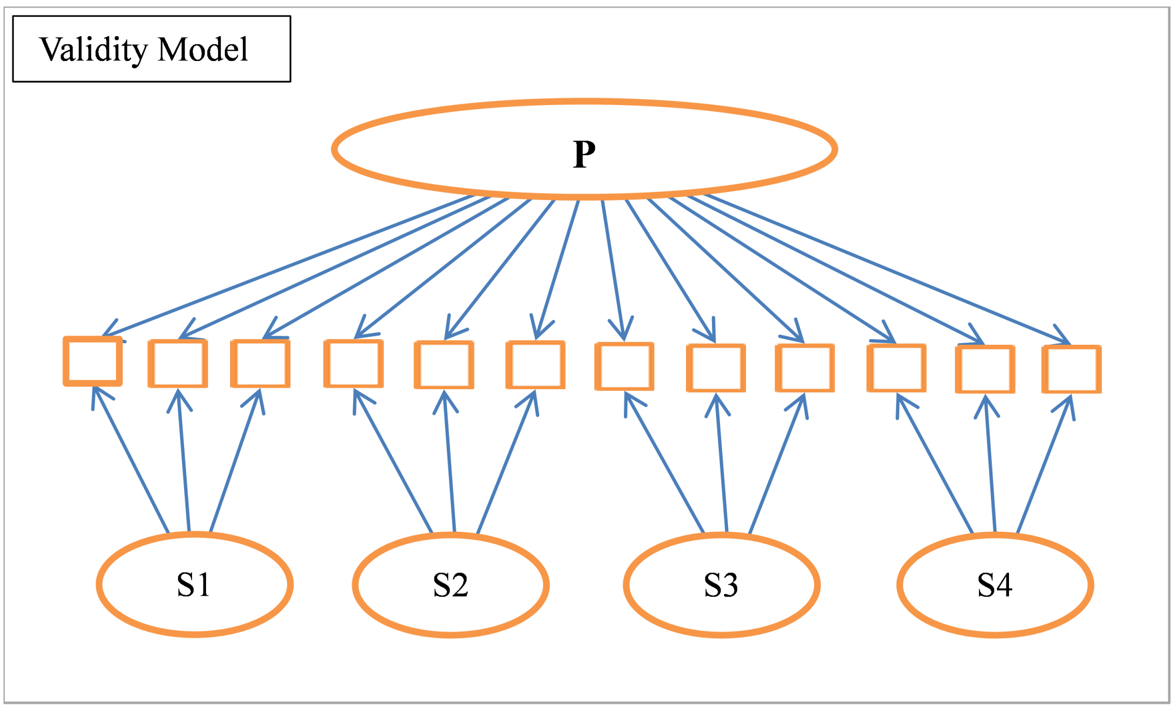

Figure 3.1: The Bifactor Model Without the Orthogonality Violation

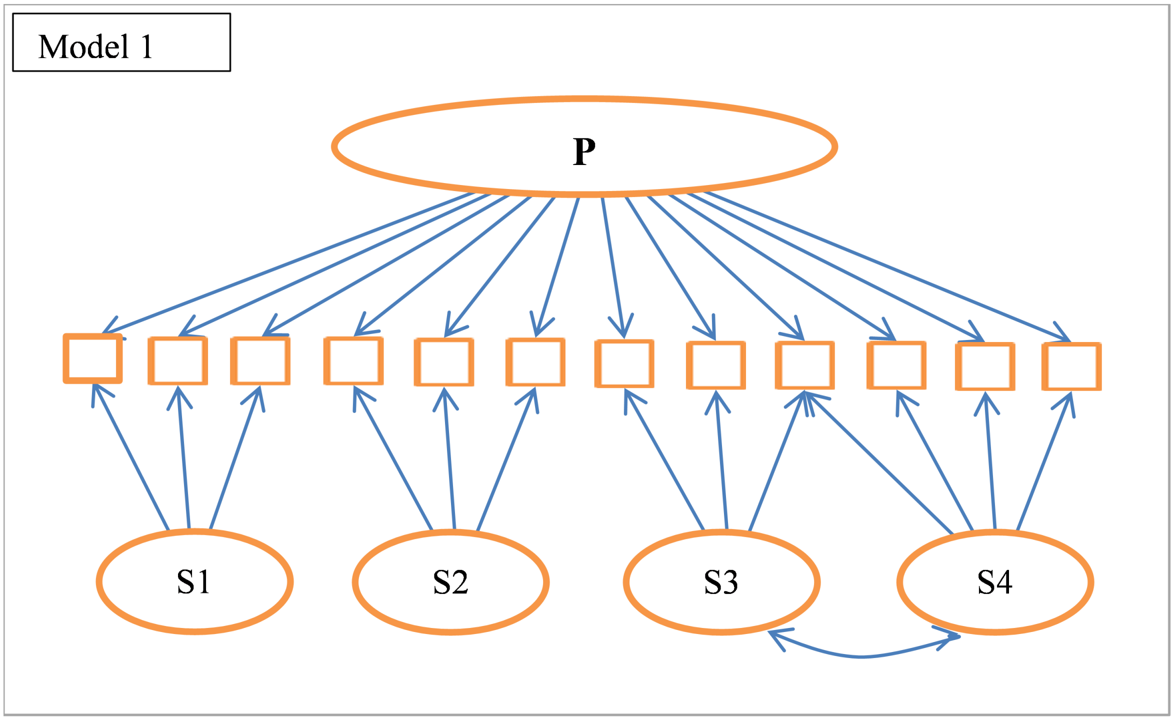

Figure 3.2: The Bifactor Model With the Orthogonality Violation Between two Subdomains

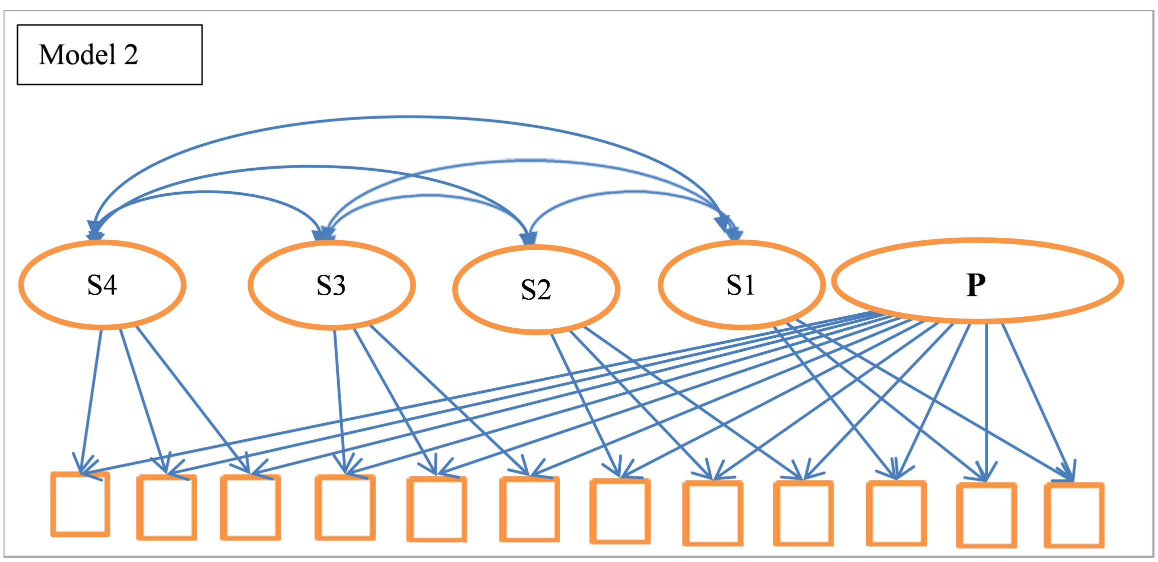

Figure 3.3: The Bifactor Model With the Orthogonality Violation Among all Subdomains

3.1.1 Validity Model: Accuracy of the Bifactor Program

To test the accuracy of the bifactor program written in *R* software, data without orthogonality violations among all factors were generated (see Figure 3.1). The simulated data had five dimensions with one primary and four subdomains, 80 items, and 5,000 participants. The item and true person parameters were saved in an Excel file, and then a bifactor analysis was conducted to obtain the parameter estimates. The estimated parameters were then compared with the true parameters to test the estimate recovery. Because the data met the orthogonal assumption of the bifactor item response theory model, it was hypothesized that parameter estimates of a correctly written program should accurately recover the true item and person parameters. If there were any differences among the parameter estimates due to random error, the differences were expected to be trivial. Table 3.1 was the correlation matrix specified in the simulation study.

Table 3.1: The Bifactor Model Without Orthogonality Violation

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.000	1.000		
S3	0.000	0.000	0.000	1.000	
S4	0.000	0.000	0.000	0.000	1.000

Table 3.2 indicates the intercept and the discrimination bias of the validity model. Results indicated that the intercept and the discrimination parameters could be accurately recovered because the mean bias between the estimated and the true item parameter estimates were centered around zero ($\overline{X}_{Int} = 0.040$; $\overline{X}_p = 0.007$; $\overline{X}_{S1} = -0.017$; $\overline{X}_{S2} = -0.008$; $\overline{X}_{S3} = 0.012$; $\overline{X}_{S4} = -0.008$), and the standard deviation statistics revealed a small variability in the bias scores between the estimated and the true parameters ($\sigma_{Int} = 0.033$; $\sigma_p = 0.021$; $\sigma_{S1} = 0.020$; $\sigma_{S2} = 0.025$; $\sigma_{S3} = 0.009$; $\sigma_{S4} = 0.019$). There were a few outliers indicated by the minimum and the maximum statistics, but most of the bias scores were small. Figure 3.4 is the graphical display to help visualize the distribution of the item parameter bias. As shown in the figure, the bias between the estimated and the true parameters was trivial. The hypothesis that a correctly written program

could recover the true parameters was accepted. The bifactor program written in *R* was correct.

Table 3.2: Discrimination and Intercept Parameter Estimate Bias of the Validity Model

	$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Int	0.040	0.033	-0.056	0.022	0.036	0.053	0.195
P	0.007	0.021	-0.060	-0.002	0.011	0.021	0.043
S1	-0.017	0.020	-0.066	-0.029	-0.012	-0.001	0.007
S2	-0.008	0.025	-0.071	-0.017	0.003	0.007	0.020
S3	0.012	0.009	-0.008	0.007	0.015	0.018	0.028
S4	-0.008	0.019	-0.053	-0.013	-0.001	0.004	0.013

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st}quartile, Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

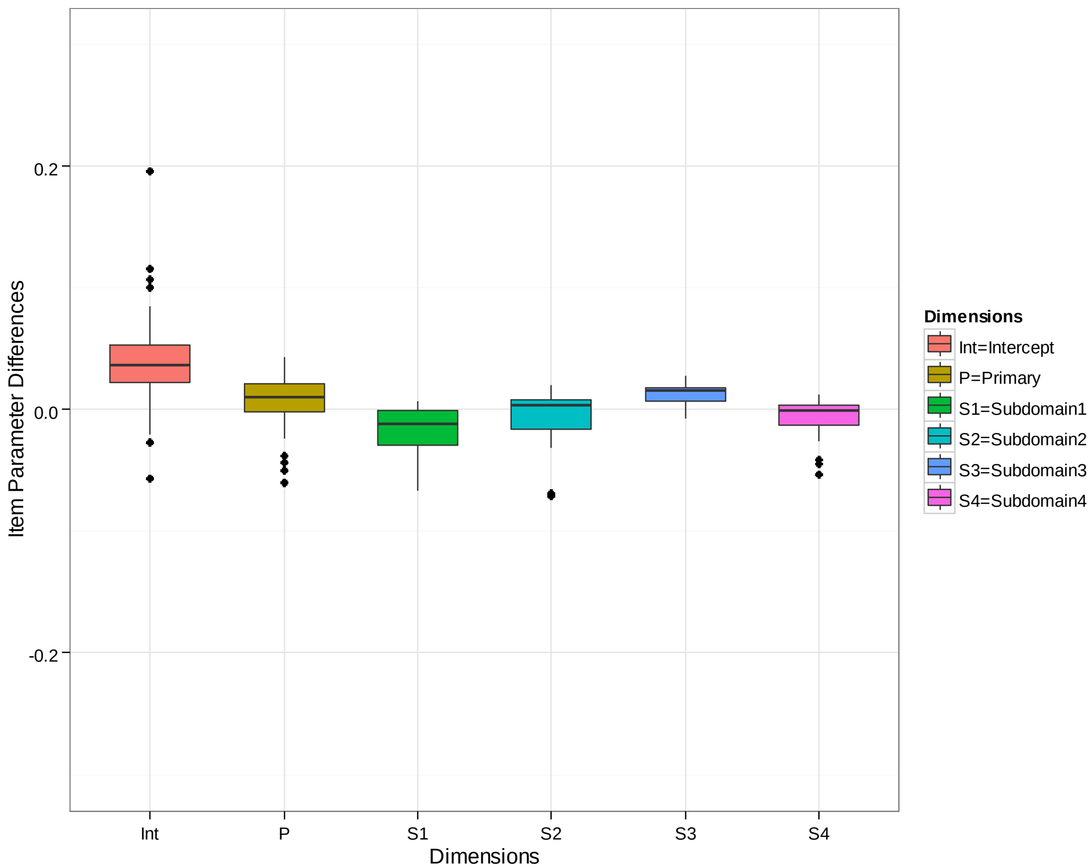

Figure 3.4: Discrimination and Intercept Parameter Estimate Bias of the Validity Model

3.1.2 Model 1: Levels of Orthogonality Violation Between two Subdomains

Four levels of orthogonality violation were designed between the third and fourth subdomains (see Figure 3.2). The four levels of orthogonality violation were: $r_{3,4}$ =0.1 (trivial violation), $r_{3,4}$= 0.2 (small violation), $r_{3,4}$= 0.4 (medium violation), and $r_{3,4}$= 0.6 (high violation). The orthogonality violations among all other dimensions were constrained to zero because the research interest was in whether the violations between two dimensions could affect parameter estimates of other dimensions. Constraining the correlations among all other dimensions to zero can reduce the confounding factors in parameter estimates. Tables 3.3 through 3.6 present the trivial, small, medium, and high orthogonality violations in data simulation.

Table 3.3: The Bifactor Model With Trivial Orthogonality Violation (Model 11)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.000	1.000		
S3	0.000	0.000	0.000	1.000	
S4	0.000	0.000	0.000	0.100	1.000

Table 3.4: The Bifactor Model With Small Orthogonality Violation (Model 12)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.000	1.000		
S3	0.000	0.000	0.000	1.000	
S4	0.000	0.000	0.000	0.200	1.000

Table 3.5: The Bifactor Model With Medium Orthogonality Violation (Model 13)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.000	1.000		
S3	0.000	0.000	0.000	1.000	
S4	0.000	0.000	0.000	0.400	1.000

Table 3.6: The Bifactor Model With High Orthogonality Violation (Model 14)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.000	1.000		
S3	0.000	0.000	0.000	1.000	
S4	0.000	0.000	0.000	0.600	1.000

3.1.3 Model 2: Levels of Orthogonality Violation Among all Subdomains

Results obtained from orthogonality violation between two subdomains can indicate whether a parameter estimate bias exists only between the two involved dimensions or among other dimensions as well. However, it is also possible that the orthogonality violation exists in all subdomains. Thus, the violation among all subdomains was designed to examine the parameter estimates bias. There were also four levels of orthogonality violation among all the subdomains. They were: $r_{2:4} \leq 0.1$(trivial violation), $0.1 < r_{2:4} \leq 0.3$ (small violation), $0.3 < r_{2:4} \leq 0.5$ (medium violation), and $r_{2:4} > 0.5$ (high violation). Tables 3.7 through 3.10 list correlation matrices of trivial, small, medium, and high orthogonality violations among all of the subdomain dimensions, respectively.

Table 3.7: The Bifactor Model With Trivial Orthogonality Violation (Model 21)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.020	1.000		
S3	0.000	0.040	0.080	1.000	
S4	0.000	0.060	0.090	0.100	1.000

Table 3.8: The Bifactor Model With Small Orthogonality Violation (Model 22)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.130	1.000		
S3	0.000	0.160	0.230	1.000	
S4	0.000	0.190	0.260	0.200	1.000

Table 3.9: The Bifactor Model With Medium Orthogonality Violation (Model 23)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.330	1.000		
S3	0.000	0.360	0.430	1.000	
S4	0.000	0.390	0.460	0.400	1.000

Table 3.10: The Bifactor Model With High Orthogonality Violation (Model 24)

	P	S1	S2	S3	S4
P	1.000				
S1	0.000	1.000			
S2	0.000	0.530	1.000		
S3	0.000	0.560	0.630	1.000	
S4	0.000	0.590	0.660	0.600	1.000

3.1.4 Number of Items

The total number of items were set to 40, 60, and 80 because most educational and psychological assessment instruments contain between 40 and 80 items. Because the latent factors for this study were fixed at five dimensions, and all subdomains were designed to have an equivalent number of items, there were 10 items for each subdomain for the 40-item design, 15 for the 60-item design, and 20 for the 80-item design. All items developed for the subdomains were intended to measure the same primary factor.

3.1.5 Sample Size

Models based on item response theory generally necessitate a large sample size. How large the sample size should be depends on the model selected. Generally speaking, the more complicated the model, the larger sample size it requires. For instance, two-parameter logistic models need larger samples than one-parameter logistic models, and similarly, three-parameter logistic models need larger samples than two-parameter logistic models. Because a bifactor model is more complex in structure and estimation procedure, a sample size of 5,000 participants was generated to guarantee the sample size was large enough for the bifactor analysis.

3.1.6 Summary of Simulation Design

The bifactor model in this study was designed to have five dimensions with one primary and four subdomains. The validity model is an ideal bifactor model containing no orthogonality violations among latent constructs. It was used to test the accuracy of a bifactor program. Model 1 refers to the orthogonality violation between two subdomains. It was used to test whether a violation among two dimensions could affect parameter estimates of the involved dimensions or others. Model 2 refers to orthogonality violation among all subdomains. It was used to test the parameter bias when there existed violations among all subdomains.There were four levels of orthogonality violation for Model 1 and Model 2. These orthogonality violations were trivial, small, medium, and large.

Fixed factors in the simulation included the number of the dimensions (five dimensions, consisting of a general dimension and four subdomains), the number of items in each subdomain, and sample size. Manipulated factors included levels of orthogonality violation and total number of items. The total number of items were designed to be 40, 60, and 80, and the sample size was 5,000. There were a total of 24 cases: two models, four levels of violation, and three numbers of item ($2 \times 4 \times 3$). Table 3.11 is the summary of the research design.

Table 3.11: Research Design of the Simulation Study

		Violation	No. of Items
Model 1	Model 11	$r_{3,4} = 0.1$	40-60-80
	Model 12	$r_{3,4} = 0.2$	40-60-80
	Model 13	$r_{3,4} = 0.4$	40-60-80
	Model 14	$r_{3,4} = 0.6$	40-60-80
Model 2	Model 21	$r_{2:4} \leq 0.1$	40-60-80
	Model 22	$0.1 < r_{2:4} \leq 0.3$	40-60-80
	Model23	$0.3 < r_{2:4} \leq 0.5$	40-60-80
	Model24	$r_{2:4} > 0.5$	40-60-80

3.2 Data Generation

To generate the two-parameter bifactor data, the distribution of the discrimination parameters (a), the difficulty parameters (b), and the ability parameters (θ) need to be specified. The discrimination parameters (a) were specified to follow uniform distribution ranging from 0.2 to 2.0, the difficulty parameters (b) were set to follow random normal distribution, and the theta parameters (θ) were set to follow multivariate normal distribution (see Table 3.12). The distribution specification for discrimination, difficulty, and ability indices were the same across all 24 cases. For every generated parameter, a random seed was set up to make sure the exact same parameter would be generated later for examination. Since the probability of a correct response is normally expressed in terms of discrimination and intercept parameters, difficulty parameters are transformed to be intercept parameters by the formula below:

$$d = -b\sqrt{a_1^2 + a_s^2}$$

Table 3.12: Item and Person Parameter Specification of the Simulation Study

		Distribution
Item and Person Parameters	a	Uniform distribution: (0.2, 2.0)
	b	Random normal distribution: (0, 1)
	θ	Random multivariate normal distribution: (0, 1)

With the specification of discrimination, intercept, and ability parameters, the probability of a correct response for a 2PL bifactor item response model can be computed as:

$$p(y = 1 \mid \theta_g, \theta_s) = \frac{1}{1 + exp^{\{-(d + a_g\theta_g + a_s\theta_s\}}}$$

Where

d is item intercept;

$a_g$is the item slope on the primary factor; and

$a_s$is the item slope on the specific factor.

The probability of a correct response given the two-parameter logistic bifactor model was obtained for each item, conditional on the ability (θ). These model-based probabilities indicate the percentage of correct responses (1s) if the samplings are done an infinite number of times. For instance, the model-based probability is 0.6, indicating that 60% of the responses will be 1s and 40% will be 0s for an infinite number of samplings. The current study had 200 replications for each of the 24 cases. All of the simulations were implemented by *R* 2.15.

3.3 Evaluation Criteria

Common criteria to assess the accuracy of parameter estimation over replications are bias, RMSE, and standard error of estimates (SE). They are computed by averaging each of the values over all items or ability parameter estimates across all replications.

3.3.1 Bias, RMSE, and SE

Evaluation criteria of bias, RMSE, and SE were computed in each of the 24 cases. Bias is the average difference between parameter estimates and the true parameters. It reflects the deviation of parameter estimates from the true values; the smaller the absolute bias, the more accurate the parameter estimates. RMSE indicates the overall parameter estimation accuracy. The smaller RMSE, the more accurate the estimates. SE indicates the stability of parameter estimates; the smaller the SE, the more stable the estimates. The evaluation criteria are computed as:

$$Bias(\hat{\beta}) = \frac{\sum_{r=1}^{R}(\hat{\beta}_r - \beta)}{R}$$

$$RMSE(\hat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \beta)^2}$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r - \frac{\sum_{r=1}^{R}\hat{\beta}_r}{R})^2}$$

Where

β is the true ability or item parameter for the true data generation model; and

$\hat{\beta}$is the estimate and the ability or item parameters at the r^{th} replication (r = 1, 2, 3, ..., R) from these estimation models (Li & Rupp, 2011). Evaluation criteria of bias and RMSE are further studied to examine the parameter estimate recovery.

3.3.2 ANOVA

Although statistics such as bias, RMSE, and SE describe the precision and efficiency of the parameter estimates, they cannot detect the magnitude of effects. ANOVA was performed to compare the mean RMSE of different models. The RMSE of the item and person parameter were the dependent variables, and the group variables were models with orthogonality violations among different subdomains, numbers of items, and levels of orthogonality violation .

3.4 Summary of Procedures

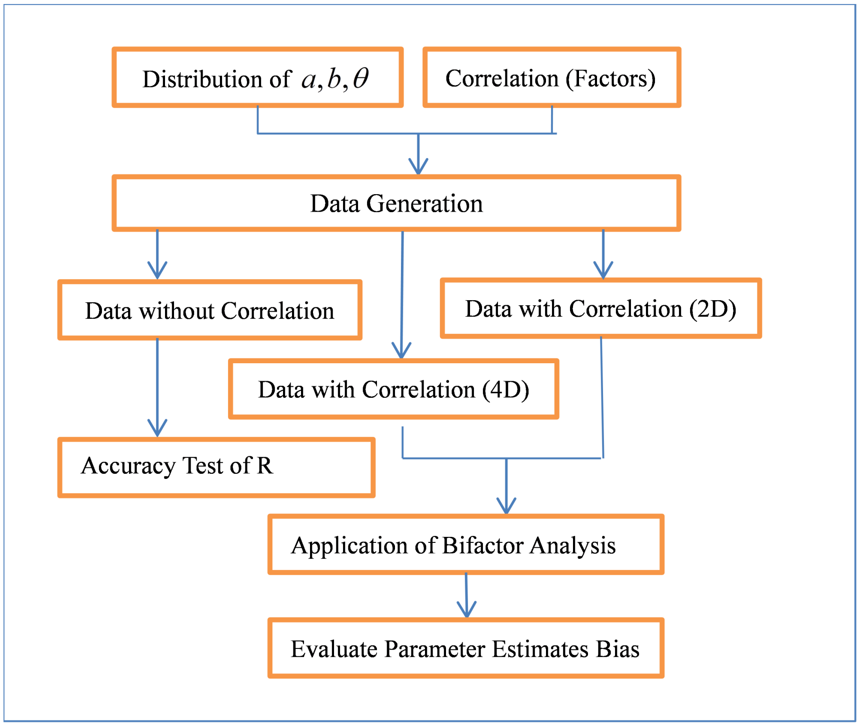

Figure 3.5: Simulation Procedures of the Study

To conclude, the procedure of data simulation can be briefly described by three phrases. First, the distributions of the item and person parameters and the correlation matrix among subdomains was specified before the actual data simulation. With the specification of the true parameters, a series of data sets under various conditions were generated. In the current study, the manipulated factors included four levels of orthogonality violation between and among subdomains and three different total numbers of items. There were 24 conditions for the simulated data. Each condition had 200 replications to minimize the sample variance and increase the power to detect the effects of interest.

Second, the validity of the bifactor program was tested. Data without violations (i.e., uncorrelated factors) were simulated to serve this purpose. The true person and item parameters were saved in an Excel file, and then the bifactor analysis based on the simulated data was conducted to test the recovery of the parameter estimates. Since these data met the orthogonal assumption of the bifactor model, parameter estimates were hypothesized to be approximately identical with true parameters.

Finally, the bifactor analysis was applied to the simulated data with violations between and among subdomains. The mean of parameter estimates was computed at over 200 replications. ANOVA was performed to compare the true and the averaged simulated parameters to examine the magnitude of effects. Commonly used criteria such as bias, RMSE, and SE, were applied to assess the accuracy of parameter estimates. They were computed by averaging each of the values over all items or ability parameter estimates. The bifactor analysis of all simulations was implemented with Mplus 7.0 software.

3.5 Empirical Study with Real Data

The real data set is the general math assessment from the Center for Educational Testing and Evaluation (CETE). The Kansas Mathematics Assessment is designed to assess all grades. The eighth-grade math assessment results from form 797 were selected to conduct the empirical study. All data were from the 2009 Kansas Computerized Assessment. The item format was multiple-choice with one key and three distractors. The math ability was measured by four subdomains: number, data, geometry, and algebra factors. The total number of items was 86: 28 for number, 23 for algebra, 18 for geometry, and 17 for data dimension. The structure of the math assessment is described in Figure 3.6:

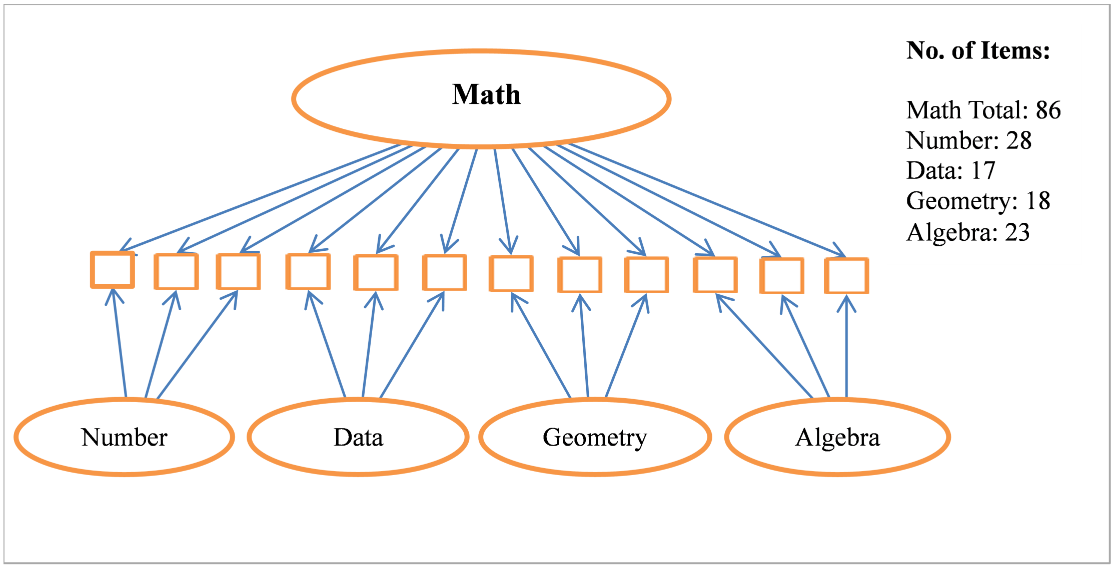

Figure 3.6: Structure of the Real Data

This data is appropriate for the bifactor structure with one primary factor and four subdomains. The primary factor is the math ability, while the four subdomains are number, data, geometry, and algebra. The number factor is further measured by three indicators: number systems and their properties, number sense, and computation; the data factor by statistics and probability; the geometry factor by geometric figures and their properties and geometry from an algebraic perspective; and the algebra factor by functions, variables, equations, inequalities, and models. Because it is difficult to write items that can measure one indicator only, there may exist some cross loadings among the factors.

3.5.1 Sample

The total sample size of the eighth-grade math was 14,457. The data included students who received free/reduced lunch and various accommodations, as well as students of different races, students from special education, and students for whom English was a second language. The research focus was on the latent bifactor structure rather than group differences; therefore, all of the aforementioned demographic variables were included in the data analysis. Because a sample size of 5,000 participants should be large enough to conduct the bifactor analysis, approximately 35%

of the total sample using SPSS software was randomly selected to decrease the implementation time and also match the simulation sample size. The final total sample size was 5,023.

3.5.2 Analysis

A bifactor analysis with one primary dimension and four subdomains was implemented with the sample by Mplus. The estimation method used was the robust maximum likelihood with a Monte Carlo integration, which was exactly the same as the ones used for simulation studies. The item and person parameters were then obtained, and a target oblique rotation method was subsequently conducted to explore the orthogonality violation among the four subdomains. The magnitude of the orthogonality violation was reported, and possible bias was then discussed.

A target oblique rotation can relax the orthogonality restriction of a bifactor model. The target matrix can specify the bifactor structure with one primary factor and a number of related subdomains, and the oblique rotation can relax the orthogonality restriction. Factors are allowed to be correlated with the oblique rotation. Two software packages are available for specifying the target oblique rotation: Comprehensive Exploratory Factor Analysis (CEFA) (Browne, Cudeck, Tateneni, & Mels, 2004) and MPLUS (Asparouhov & Muthen, 2009). Currently, the target rotation was applied to conduct exploratory bifactor analysis and was used as a comparison model to investigate the robustedness of unidimensional item response theory parameter estimates bias when local independence was violated (Reise, Moore, & Maydeu-Olivares, 2011; Reise et al., 2010). Simulation studies of Reise et al. (2011) indicated that target rotation could accurately recover the factor loadings when the orthogonality assumption of the bifactor model was met.

Chapter 4

Results

Results are presented in the order of the specific research questions. The first section, Parameter Estimate Bias of Model 1, describes the item and person parameter bias when the orthogonality violations exist only between two subdomains. The second section, Parameter Estimate Bias of Model 2, describes the parameter bias when the orthogonality violations exist among all subdomains. The results are arranged in a framework of item parameter estimate bias and then person parameter bias across different numbers of items and levels of orthogonality violation. The effects of numbers of items and models on item and person parameter estimates are illustrated in the third section. The mean bias and the mean RMSE are described, followed by the effects of numbers of items and models on the parameter estimates. A further analysis of specific models and dimensions on parameter estimates is presented in the last section. Similarly, the mean bias and the mean RMSE are reported first, followed by the effects of specific models and dimensions.

The software packages used in this study were Mplus 7.0 (Muthén & Muthén, 1998-2012) , *R* 2.15 (Venables, Smith, & Team, 2002), and SPSS 20.0 (Statistics, 2010). Mplus software was used for the bifactor analysis. Most of the result integration was conducted using *R* 2.15. The *R* packages used in this study include "MplusAutomation" (Hallquist, 2011) for the parameter extraction; "ggplot2" (Wickham, 2008) for the graphical display; "doBy" (Højsgaard, 2012) for the descriptive statistics such as mean, median, and quartiles; "kobe" (Kell, 2012) for multiple graph display; and "lsr" (Navarro, 2013) for post hoc analysis of ANOVA results. The effect of models, numbers of items, and dimensions on parameter estimates were conducted by SPSS.

4.1 Parameter Estimate Bias of Model 1

This section describes the parameter estimate bias when orthogonality violations exist only between the third and fourth subdomains. Levels of orthogonality violation are labeled from the least to the most severe as Model 11 ($r_{3,4}$ =0.1), Model 12 ($r_{3,4}$= 0.2), Model 13 ($r_{3,4}$= 0.4), and Model 14 ($r_{3,4}$= 0.6). The bias between the estimated and the true parameters are tabled in terms of numbers of items, specific models, and dimensions. Descriptive analysis such as the mean, the standard deviation, and the quartiles of the bias are reported to compare the recovery between the estimated and the true parameters across different conditions.

4.1.1 Item Parameter Estimate Bias of 40 Items

To examine the recovery of the item parameter estimates, the bias was computed by subtracting the true parameter estimates from the estimated parameter estimates. Table 4.1 summarizes the item parameter estimate bias of 40 items. As indicated by the table, intercept parameters could be accurately recovered because the mean bias scores of all specific models were centered around zero ($\overline{X}_{Model11} = 0.005$, $\overline{X}_{Model12} = -0.001$, $\overline{X}_{Model13} = 0.010$, $\overline{X}_{Model14} = -0.003$), and the variability of the intercept parameter bias was also small ($\sigma_{Model11} = 0.049$, $\sigma_{Model12} = 0.032$, $\sigma_{Model13} = 0.032$, $\sigma_{Model14} = -0.052$). There were a few outliers for each specific model (e.g. $Max_{Model11} = 0.200$, $Min_{Model11} = -0.187$), but for the most part the bias was very small. The small deviation suggested that the estimated intercept parameters did not differ much from the true intercept parameters; thus, levels of orthogonality violation had little influence on intercept parameter estimates.

Additionally, the orthogonality violation between the third and fourth dimensions did not have much effect on the discrimination parameter estimates of the first and second subdomains.

The average discrimination parameter bias of the first and second subdomains was small ($\overline{X}_{S1\&Model11} = -0.003$, $\overline{X}_{S1\&Model12} = -0.007$, $\overline{X}_{S1\&Model13} = -0.017$, $\overline{X}_{S1\&Model14} = -0.005$; $\overline{X}_{S2\&Model11} = -0.031$, $\overline{X}_{S2\&Model12} = 0.017$, $\overline{X}_{S2\&Model13} = 0.009$, $\overline{X}_{S2\&Model14} = 0.028$), and the standard deviation of the bias scores indicated a trivial variability (e.g., $\sigma_{S1\&Model11} = 0.018$, $\sigma_{S1\&Model12} = 0.014$, $\sigma_{S1\&Model13} = 0.023$, $\sigma_{S1\&Model14} = 0.062$). These results indicated that the orthogonality violation between two subdomains did not influence the parameter estimate accuracy of other subdomains.

However, the orthogonality violation did affect the parameter estimates of the involved dimensions. The discrimination parameters of the third and fourth subdomains demonstrated large discrepancies between the estimated and the true parameters as the orthogonality violations became large ($\overline{X}_{S3\&Model11} = -0.053$, $\overline{X}_{S3\&Model12} = -0.069$, $\overline{X}_{S3\&Model13} = -0.132$, $\overline{X}_{S3\&Model14} = -0.252$; $\overline{X}_{S4\&Model11} = -0.077$, $\overline{X}_{S4\&Model12} = -0.073$, $\overline{X}_{S4\&Model13} = -0.148$, $\overline{X}_{S4\&Model14} = -0.241$). Large orthogonality violations resulted in large bias scores between the estimated and the true parameter. The large violation affected the bias range as well. The standard deviation indicated that the more severe the orthogonality violation, the more variable the bias became (e.g., $\sigma_{S3\&Model11} = 0.016$, $\sigma_{S3\&Model12} = 0.022$, $\sigma_{S3\&Model13} = 0.045$, $\sigma_{S3\&Model14} = 0.076$). The results clearly demonstrated that violations between dimensions did influence the parameter estimate recovery of the involved dimensions.

The orthogonality violations between two dimensions also affected the parameter estimate recovery of the primary dimension. The mean bias between the estimated and the true parameters had small differences ($\overline{X}_{P\&Model11} = 0.023$, $\overline{X}_{P\&Model12} = 0.030$, $\overline{X}_{P\&Model13} = 0.034$, $\overline{X}_{P\&Model14} = 0.045$). However, the standard deviation showed that the bias range was much different among the specific models. Large violations resulted in a large variability of the bias range ($\sigma_{P\&Model11} = 0.029$, $\sigma_{P\&Model12} = 0.040$, $\sigma_{P\&Model13} = 0.072$, $\sigma_{P\&Model14} = 0.159$). The results indicated that the orthogonality violations between subdomains influenced the variance, but not the mean bias of the primary dimension.

Figure 4.1 is the graphical display of discrimination and intercept estimate bias for 40 items.

As shown by the figure, intercept parameters as well as discrimination parameters of the first and the second subdomains had the least deviance from the true parameters. The parameter estimate bias of the primary dimension, the third, and the fourth sudomains worsened as the violations increased. In addition, the figure illustrates that all of the parameters for the third and the fourth dimensions were underestimated because the differences between the estimated and the true were all below zero.

Table 4.1: Discrimination and Intercept Parameter Estimate Bias of 40 Items (Model 1)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 11	Int	0.005	0.049	-0.187	-0.011	0.005	0.021	0.200
	P	0.023	0.029	-0.035	-0.001	0.028	0.042	0.072
	S1	-0.003	0.018	-0.032	-0.015	-0.006	0.010	0.027
	S2	-0.031	0.022	-0.059	-0.051	-0.032	-0.012	0.002
	S3	-0.053	0.016	-0.074	-0.061	-0.058	-0.050	-0.024
	S4	-0.077	0.078	-0.220	-0.087	-0.038	-0.032	-0.015
Model 12	Int	-0.001	0.032	-0.160	-0.007	0.001	0.005	0.096
	P	0.030	0.040	-0.027	-0.005	0.018	0.063	0.107
	S1	-0.007	0.014	-0.030	-0.017	-0.004	0.004	0.013
	S2	0.017	0.017	-0.024	0.016	0.019	0.024	0.043
	S3	-0.069	0.022	-0.096	-0.089	-0.069	-0.056	-0.034
	S4	-0.073	0.050	-0.156	-0.099	-0.053	-0.041	-0.019
Model 13	Int	0.010	0.032	-0.125	-0.002	0.004	0.019	0.096
	P	0.034	0.072	-0.110	-0.025	0.023	0.087	0.211
	S1	-0.017	0.023	-0.041	-0.039	-0.017	0.001	0.018
	S2	0.009	0.024	-0.035	-0.003	0.005	0.024	0.048
	S3	-0.132	0.045	-0.201	-0.165	-0.134	-0.097	-0.070
	S4	-0.148	0.062	-0.232	-0.194	-0.156	-0.105	-0.043
Model 14	Int	-0.003	0.052	-0.193	-0.017	-0.007	0.017	0.176
	P	0.045	0.159	-0.217	-0.093	0.024	0.183	0.339
	S1	-0.005	0.062	-0.084	-0.063	0.002	0.026	0.088
	S2	0.028	0.051	-0.046	-0.008	0.027	0.054	0.111
	S3	-0.252	0.076	-0.355	-0.308	-0.264	-0.211	-0.124
	S4	-0.241	0.097	-0.400	-0.310	-0.240	-0.174	-0.084

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st}quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

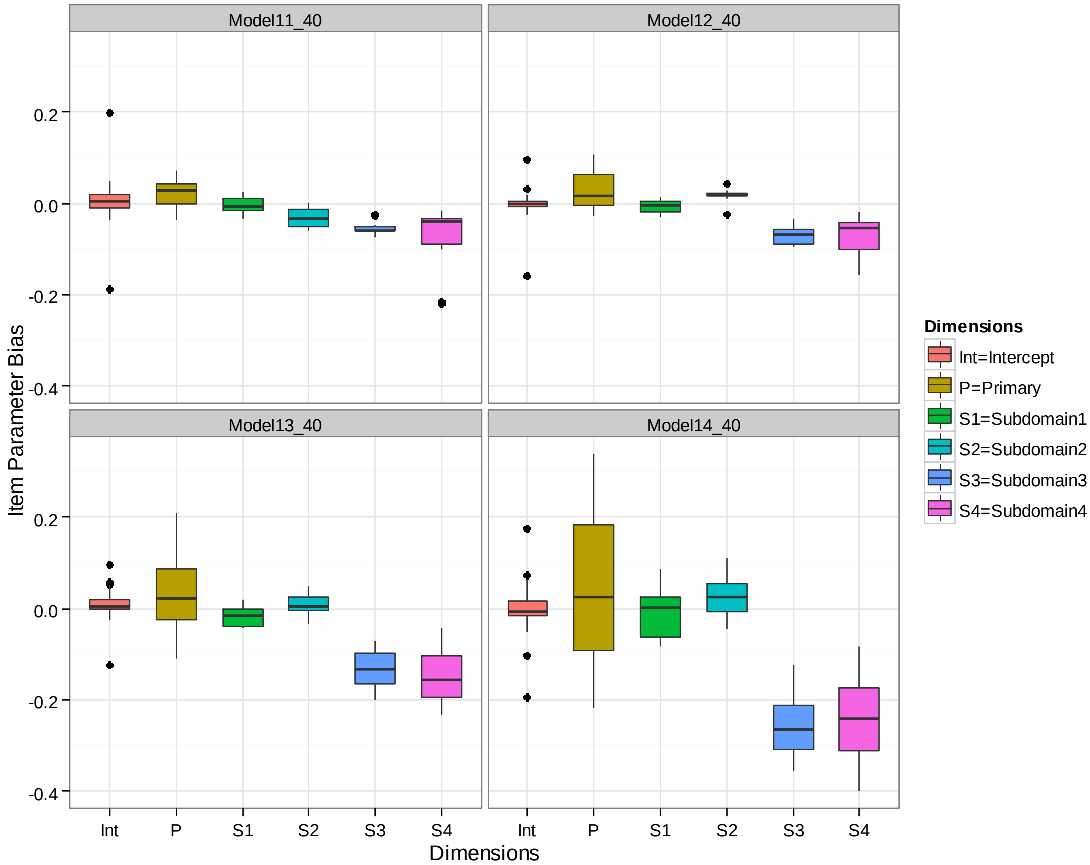

Figure 4.1: Discrimination and Intercept Parameter Estimate Bias of 40 Items (Model 1)

4.1.2 Item Parameter Estimate Bias of 60 Items

Table 4.2 reports the item parameter estimate bias of 60 items. The parameter estimate recovery of 60 items demonstrated a result similar to that of 40 items. For intercept parameters, levels of orthogonality violation could not affect the parameter estimate accuracy. The deviation between the estimated and the true intercept parameters was small because the mean bias scores of all dimensions were centered around zero ($\overline{X}_{Model11} = 0.011$, $\overline{X}_{Model12} = 0.031$, $\overline{X}_{Model13} = 0.028$, $\overline{X}_{Model14} = 0.009$), and the variability of the bias scores was also small ($\sigma_{Model11} = 0.048$, $\sigma_{Model12} = 0.053$, $\sigma_{Model13} = 0.047$, $\sigma_{Model14} = 0.043$). There were a few outliers, but overall the true intercept parameters could be recovered regardless of the orthogonality violation levels.

For the discrimination parameters, the true parameters of the first and second subdomains could be recovered because the mean bias scores were close to zero and the bias range was small (e.g., $\overline{X}_{S1\&Model11} = -0.027$, $\overline{X}_{S1\&Model12} = -0.008$, $\overline{X}_{S1\&Model13} = -0.014$, $\overline{X}_{S1\&Model14} = 0.043$;); therefore, orthogonality violations between the third and fourth sudomains could not influence the parameter accuracy of the first and second subdomains. But the orthogonality violations did influence the involved dimensions. The bias scores of the third and fourth subdomains became larger as the levels of the orthogonality violations between the subdomains became more severe (e.g.,$\overline{X}_{S3\&Model11} = -0.045, \overline{X}_{S3\&Model12} = -0.071, \overline{X}_{S3\&Model13} = -0.136, \overline{X}_{S3\&Model14} = -0.253$). The violations not only affected the mean bias but also the bias range of the involved dimensions. Large violations led to wider range of the bias distribution ($\sigma_{S4\&Model11} = 0.052$, $\sigma_{S4\&Model12} = 0.077$, $\sigma_{S4\&Model13} = 0.070$, $\sigma_{S4\&Model14} = 0.121$).

As with the results of 40 items, the violations between the third and the fourth subdomains also affected the estimates of the primary dimensions. Severe violations did not lead to large differences in the mean bias score but did result in large variability between the estimated and true parameters. There were not obvious differences in estimates between 40 and 60 items. Figure 4.2 is the graphical display of the discrimination and intercept parameter estimates of 60 items. All of the parameters of the third and fourth dimensions were underestimated because the differences were all below zero.

Table 4.2: Discrimination and Intercept Parameter Estimate Bias of 60 Items (Model 1)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 11	Int	0.011	0.048	-0.069	-0.007	0.004	0.020	0.263
	P	0.001	0.028	-0.077	-0.017	-0.002	0.023	0.049
	S1	-0.027	0.029	-0.090	-0.052	-0.020	-0.004	0.010
	S2	-0.038	0.028	-0.095	-0.056	-0.047	-0.014	0.003
	S3	-0.045	0.027	-0.110	-0.062	-0.039	-0.023	-0.013
	S4	-0.063	0.052	-0.205	-0.070	-0.056	-0.027	-0.014
Model 12	Int	0.031	0.053	-0.036	0.006	0.018	0.042	0.337
	P	0.024	0.036	-0.071	-0.001	0.018	0.043	0.098
	S1	-0.008	0.031	-0.060	-0.036	0.002	0.017	0.029
	S2	-0.015	0.032	-0.103	-0.025	-0.005	0.001	0.028
	S3	-0.071	0.031	-0.127	-0.093	-0.074	-0.049	-0.024
	S4	-0.099	0.077	-0.344	-0.110	-0.087	-0.054	-0.029
Model 13	Int	0.028	0.047	-0.063	0.008	0.020	0.041	0.283
	P	0.045	0.070	-0.098	-0.008	0.020	0.098	0.189
	S1	-0.014	0.038	-0.077	-0.038	0.001	0.012	0.031
	S2	-0.017	0.038	-0.072	-0.043	-0.013	-0.002	0.063
	S3	-0.136	0.051	-0.232	-0.172	-0.125	-0.104	-0.045
	S4	-0.153	0.070	-0.325	-0.190	-0.149	-0.105	-0.055
Model 14	Int	0.009	0.043	-0.082	-0.011	0.003	0.017	0.183
	P	0.062	0.169	-0.238	-0.070	0.004	0.198	0.369
	S1	0.043	0.037	-0.039	0.022	0.051	0.068	0.087
	S2	0.049	0.054	-0.042	0.013	0.054	0.089	0.159
	S3	-0.253	0.096	-0.435	-0.313	-0.239	-0.190	-0.084
	S4	-0.286	0.121	-0.608	-0.337	-0.261	-0.214	-0.116

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

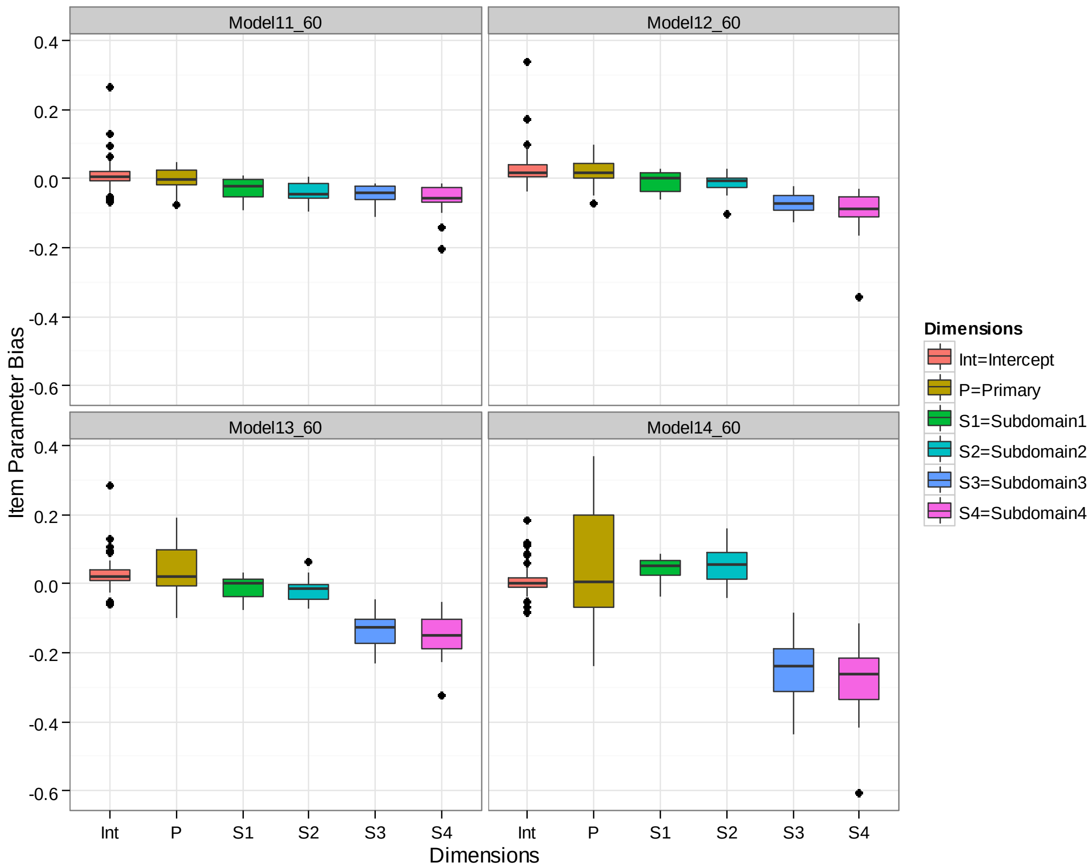

Figure 4.2: Discrimination and Intercept Parameter Estimate Bias of 60 Items (Model 1)

4.1.3 Item Parameter Estimate Bias of 80 Items

Table 4.3 describes the item parameter bias of 80 items. The results showed a trend similar to the those of 40 items and 60 items. Intercept parameters could be accurately recovered; discrimination parameters of the first and second subdomains could be recovered but not the ones of the third and fourth subdomains. The violations affected only the parameters of the involved dimensions and the primary dimension, not other subdomains. The larger the orthogonality violation, the larger the bias between the estimated and the true parameters.

However, the mean bias for the third and the fourth dimensions of 80 items was smaller than the corresponding bias of 40 and 60 items. Taking Model 14 as an example, the mean bias of 80 items was smaller than the mean bias of 40 and 60 items ($\overline{X}_{S3\&40} = -0.252$, $\overline{X}_{S4\&40} = 0.241$;

$\overline{X}_{S3\&60} = -0.253$, $\overline{X}_{S3\&60} = -0.286$; $\overline{X}_{S3\&80} = -0.152$, $\overline{X}_{S3\&60} = -0.169$), which indicated that the number of items could improve parameter estimate accuracy. In addition, the outlier values of the bias scores were smaller than the models with 40 and 60 items. the largest outliers of models with 40 items were -0.400 and 0.338. For 60 items it was -0.689 and 0.369, while for 80 items the largest outliers were -0.273 and 0.221. This result also demonstrated that the number of items could improve parameter accuracy. A further analysis is needed to examine whether the number of items is a significant predictor of parameter accuracy. Figure 4.3 is the graphical display of discrimination and intercept parameter estimates of 80 items. All parameters of the third and fourth dimensions were underestimated because all bias scores were below zero.

Table 4.3: Discrimination and Intercept Parameter Estimate Bias of 80 Items (Model 1)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 11	Int	-0.005	0.031	-0.069	-0.019	-0.008	0.008	0.093
	P	0.038	0.014	0.013	0.028	0.035	0.045	0.080
	S1	-0.034	0.023	-0.069	-0.055	-0.033	-0.011	-0.002
	S2	-0.025	0.032	-0.093	-0.048	-0.017	0.002	0.012
	S3	-0.029	0.014	-0.059	-0.038	-0.028	-0.019	-0.007
	S4	-0.064	0.034	-0.124	-0.090	-0.061	-0.036	-0.013
Model 12	Int	-0.008	0.036	-0.107	-0.025	-0.004	0.008	0.117
	P	0.048	0.033	-0.030	0.023	0.057	0.070	0.102
	S1	-0.022	0.025	-0.072	-0.043	-0.020	0.002	0.009
	S2	-0.011	0.044	-0.120	-0.021	0.005	0.017	0.033
	S3	-0.048	0.020	-0.094	-0.059	-0.044	-0.037	-0.008
	S4	-0.072	0.035	-0.141	-0.096	-0.065	-0.047	-0.021
Model 13	Int	-0.003	0.044	-0.092	-0.029	-0.005	0.014	0.172
	P	0.043	0.079	-0.092	-0.024	0.017	0.117	0.184
	S1	-0.004	0.026	-0.059	-0.022	-0.008	0.020	0.039
	S2	-0.002	0.047	-0.112	-0.020	0.010	0.032	0.046
	S3	-0.094	0.034	-0.161	-0.112	-0.086	-0.072	-0.023
	S4	-0.114	0.042	-0.183	-0.139	-0.113	-0.085	-0.026
Model 14	Int	0.008	0.033	-0.112	-0.003	0.012	0.026	0.125
	P	0.040	0.102	-0.100	-0.047	0.013	0.141	0.221
	S1	0.028	0.025	-0.021	0.005	0.026	0.047	0.067
	S2	0.004	0.055	-0.133	-0.017	0.027	0.039	0.051
	S3	-0.152	0.062	-0.273	-0.186	-0.150	-0.101	-0.044
	S4	-0.169	0.059	-0.270	-0.206	-0.162	-0.125	-0.056

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st}quartile,
Q2 = 2^{nd} quartile,Q3 = 3^{rd} quartile

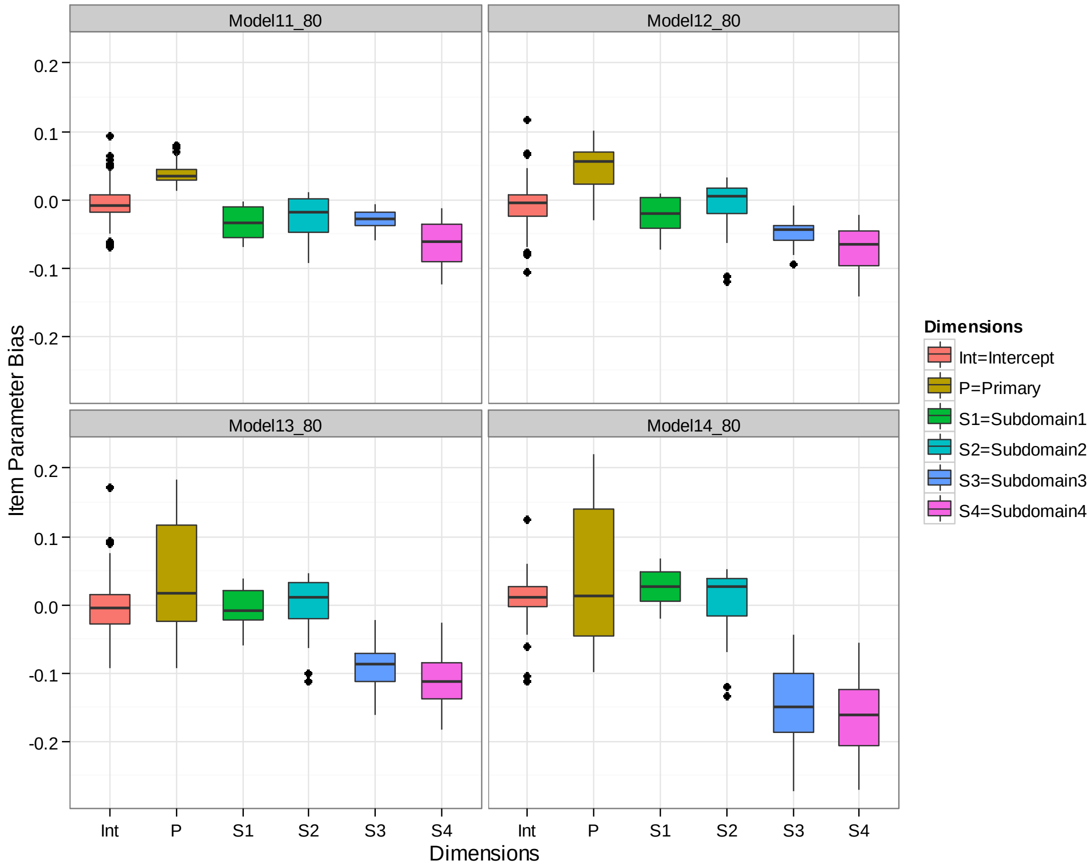

Figure 4.3: Discrimination and Intercept Parameter Estimate Bias of 80 Items (Model 1)

4.1.4 Person Parameter Estimate Bias of 40 Items

Table 4.4 reports the person parameter bias of 40 items. As shown by the table, the estimated theta parameters were much different from the true parameters across all of the specific models because the bias scores were widely spread out. Fifty percent of the bias scores were between -0.413 and 0.449 (e.g., $Q1_{P\&Model11} = -0.236$, $Q2_{P\&Model11} = -0.010$; $Q3_{P\&Model11} = 0.221$; $Q1_{S1\&Model11} = -0.199$, $Q2_{S1\&Model11} = -0.033$; $Q3_{S1\&Model11} = 0.266$), while the other 50% of the bias scores could be as large as -3.850 and 2.570. Obviously, specific models (i.e., levels of orthogonality violation) had no significant effect on the accuracy of theta parameter estimates.

The distribution of the bias scores was approximately symmetric with the mean bias scores centered around zero across all dimensions of specific models (e.g., $\overline{X}_{P\&Model11} = -0.005$,

$\overline{X}_{S1\&Model11} = 0.013$; $\overline{X}_{S2\&Model11} = -0.026$, $\overline{X}_{S3\&Model11} = 0.014$; $\overline{X}_{S4\&Model11} = 0.020$). Of all the specific models, the standard deviation of the bias scores for the primary dimension was the smallest and the standard deviation of the third subdomain was the largest ($\sigma_{P\&Model11} = 0.347$; $\sigma_{S1\&Model11} = 0.414$, $\sigma_{S2\&Model11} = 0.462$; $\sigma_{S3\&Model11} = 0.595$, $\sigma_{S4\&Model11} = 0.481$). Because the variability of the primary dimensions was the smallest of all the dimensions, the parameter estimates were relatively trustworthy.

Figure 4.4 is the graphical display of the theta parameter estimates of 40 items. The figure indicated that the mean bias of all dimensions across all specific models was approximiately the same, however the range of the bias scores was widespread. Of all dimensions, the parameter estimates of the primary dimensions were less distorted, and the parameter estimates of the third subdomain were most distorted. Overall, the theta parameter estimates were much different from the true theta parameters.

Table 4.4: Theta Parameter Estimate Bias of 40 Items (Model 1)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 11	P	-0.005	0.347	-1.751	-0.236	-0.010	0.221	1.805
	S1	0.013	0.414	-2.108	-0.199	0.033	0.266	1.769
	S2	-0.026	0.462	-2.566	-0.247	0.036	0.267	1.518
	S3	0.014	0.595	-3.040	-0.344	0.047	0.408	2.388
	S4	0.020	0.481	-1.962	-0.290	0.018	0.325	2.087
Model 12	P	-0.001	0.350	-1.523	-0.229	-0.013	0.215	1.905
	S1	-0.005	0.424	-2.895	-0.218	0.041	0.246	2.033
	S2	-0.002	0.469	-2.846	-0.236	0.054	0.298	2.196
	S3	0.013	0.591	-2.698	-0.344	0.052	0.408	2.570
	S4	-0.005	0.493	-3.850	-0.316	-0.013	0.311	2.277
Model 13	P	-0.001	0.365	-1.390	-0.239	-0.014	0.233	1.503
	S1	0.029	0.406	-2.100	-0.176	0.063	0.274	1.916
	S2	-0.003	0.462	-3.103	-0.230	0.058	0.300	1.661
	S3	-0.003	0.606	-2.858	-0.381	0.042	0.408	2.295
	S4	0.011	0.494	-1.897	-0.309	0.009	0.334	2.325
Model 14	P	-0.009	0.411	-1.455	-0.285	-0.011	0.262	1.832
	S1	-0.024	0.424	-2.789	-0.244	0.008	0.241	1.601
	S2	0.016	0.480	-2.625	-0.231	0.066	0.328	1.580
	S3	-0.001	0.674	-3.139	-0.413	0.053	0.449	2.307
	S4	0.001	0.548	-2.364	-0.352	0.003	0.357	2.414

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile, Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

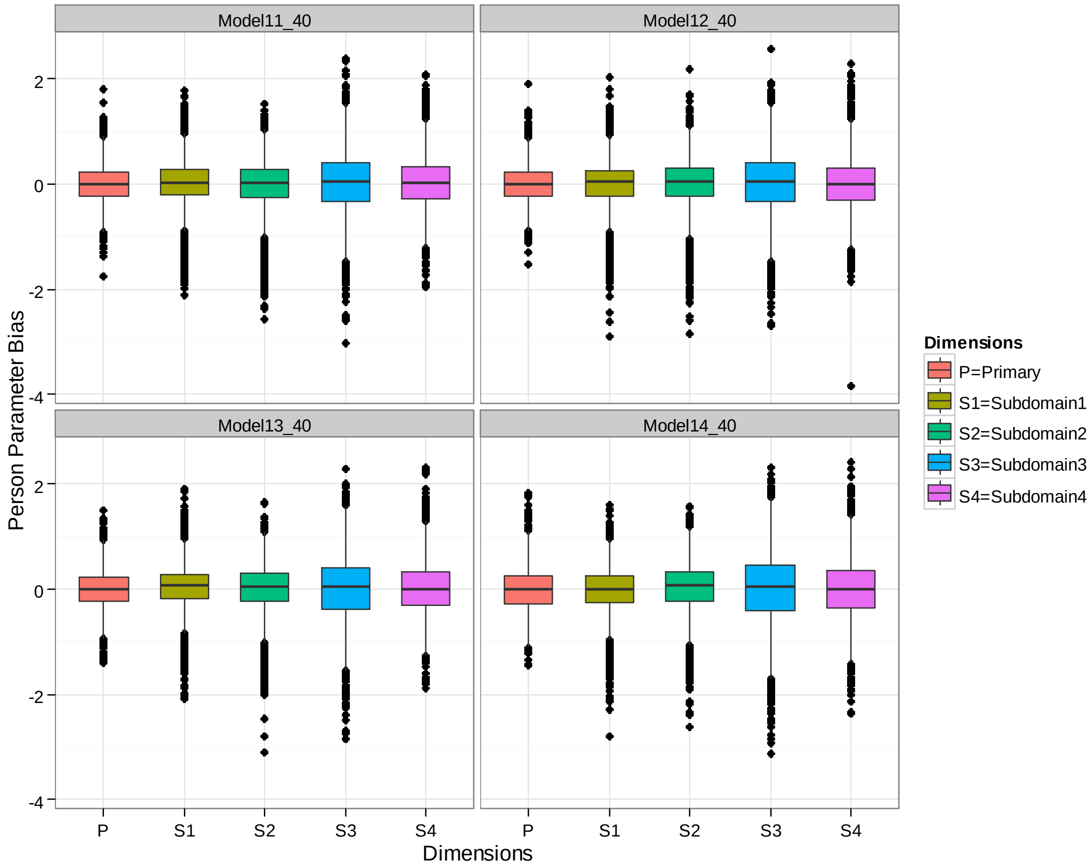

Figure 4.4: Theta Parameter Estimate Bias of 40 Items (Model 1)

4.1.5 Person Parameter Estimate Bias of 60 Items

Table 4.5 reports the theta parameter bias of 60 items. The bias distribution of each specific model was similar to that of 40-item estimates. Overall the bias scores were widely spread out. Fifty percent of the theta estimates were between -0.350 and 0.324, and the other 50% could be as large as -3.136 and 2.718. Specific models had no influence on the theta parameter estimates. The distribution of the bias scores was approximately symmetric with a mean score centered around zero. The standard deviations of the bias scores for the primary dimension were the smallest across all specific models; therefore, the parameter estimates were the most trustworthy compared with estimates of other dimensions.

The range of the bias scores of the 60 items was a little small compared with the range of

40 items. The bias range of 40 items was between -3.850 and 2.570, while the bias range of 80 items was between -3.136 and 2.718. Among all the dimensions, the estimates of the third dimension of the 40 items were the least trustworthy because the standard deviations were the largest, while in the 60-item model, the estimates of the fourth dimension were the least trustworthy (e.g., $\sigma_{P\&Model11} = 0.335$, $\sigma_{S1\&Model11} = 0.391$; $\sigma_{S2\&Model11} = 0.416$, $\sigma_{S3\&Model11} = 0.386$; $\sigma_{S4\&Model11} = 0.488$). Figure 4.5 is the graphical display of theta parameter estimates of 60 items.

Table 4.5: Theta Parameter Estimate Bias of 60 Items (Model 1)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 11	P	0.009	0.335	-1.393	-0.201	0.011	0.231	1.429
	S1	-0.016	0.391	-2.825	-0.225	0.015	0.229	1.704
	S2	-0.019	0.416	-2.232	-0.266	-0.019	0.235	1.774
	S3	0.014	0.386	-2.459	-0.208	0.021	0.233	2.261
	S4	-0.003	0.488	-2.361	-0.291	-0.016	0.284	2.250
Model 12	P	0.002	0.347	-1.283	-0.233	0.002	0.234	1.392
	S1	0.012	0.390	-1.825	-0.201	0.042	0.259	1.903
	S2	0.006	0.426	-2.543	-0.247	0.011	0.269	2.053
	S3	0.041	0.395	-2.327	-0.182	0.045	0.266	2.152
	S4	0.009	0.497	-2.378	-0.290	0.001	0.311	2.285
Model 13	P	0.012	0.369	-1.755	-0.225	0.016	0.254	1.452
	S1	0.019	0.387	-2.193	-0.191	0.047	0.264	1.571
	S2	-0.015	0.426	-2.194	-0.270	-0.015	0.238	1.931
	S3	0.008	0.402	-2.533	-0.210	0.011	0.231	2.185
	S4	0.023	0.510	-2.533	-0.279	0.008	0.324	2.255
Model 14	P	0.001	0.430	-1.850	-0.284	-0.006	0.292	1.575
	S1	0.012	0.412	-2.326	-0.230	0.031	0.276	1.757
	S2	0.007	0.451	-2.585	-0.272	0.017	0.297	2.261
	S3	0.004	0.459	-2.511	-0.236	0.013	0.255	2.718
	S4	-0.025	0.579	-3.136	-0.350	-0.026	0.300	2.461

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile,Q3 = 3^{rd} quartile

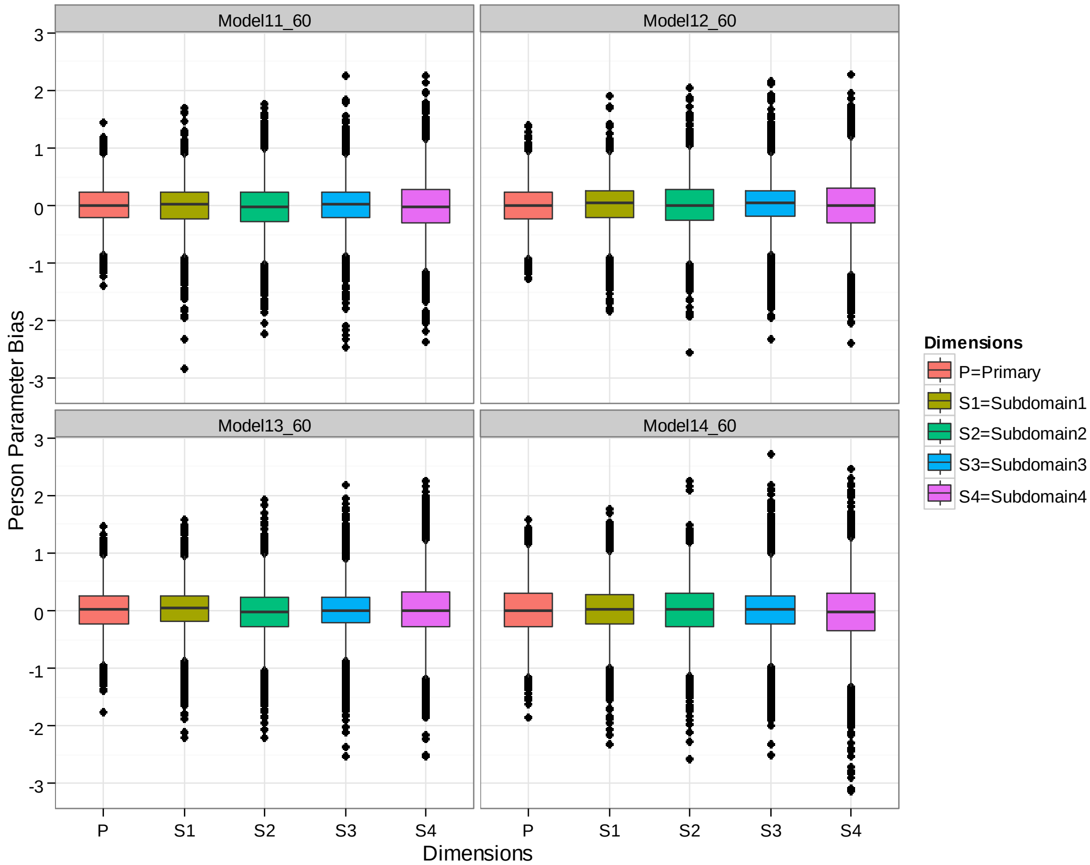

Figure 4.5: Theta Parameter Estimate Bias of 60 Items (Model 1)

4.1.6 Person Parameter Estimate Bias of 80 Items

Table 4.6 describes the theta parameter bias of 80 items. The results of parameter estimates were consistent with those of 40 and 60 items. The levels of the orthogonality violation had no influence on the parameter estimates because the bias distributions across all the specific models were approximately the same. All bias scores were centered around zero, but had wide range and many outliers (see Figure 4.6). The parameters of the first dimensions were most trustworthy because the bias range was relatively small compared with the range of other dimensions.

Taking a sample size of 80 items improved the standard deviation of the bias scores. The 50% bias scores between the 1^{st} and the 3^{rd} quartiles of 80 items were smaller than those of 40 and 60 items ($Q1_{40} = -0.413$, $Q3_{40} = 0.499$; $Q1_{60} = -0.350$, $Q3_{60} = 0.324$; $Q1_{80} = -0.258$,

$Q3_{80} = 0.218$). Figure 4.6 is the graphical display of the theta parameter of 80 items.

Table 4.6: Theta Parameter Estimate Bias of 80 Items (Model 1)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 11	P	-0.012	0.247	-1.176	-0.158	-0.010	0.142	1.228
	S1	0.020	0.308	-1.903	-0.125	0.042	0.185	1.948
	S2	0.012	0.374	-2.137	-0.193	0.001	0.218	2.004
	S3	-0.002	0.315	-1.343	-0.177	-0.021	0.153	1.861
	S4	-0.011	0.393	-2.251	-0.241	-0.030	0.204	2.082
Model 12	P	0.001	0.259	-1.412	-0.156	0.005	0.161	1.486
	S1	-0.003	0.334	-2.196	-0.149	0.019	0.169	2.551
	S2	-0.007	0.385	-2.220	-0.216	-0.019	0.205	2.156
	S3	-0.027	0.326	-3.819	-0.200	-0.040	0.135	1.891
	S4	-0.010	0.424	-3.099	-0.258	-0.040	0.203	2.071
Model 13	P	-0.006	0.269	-1.244	-0.170	-0.004	0.162	1.332
	S1	0.021	0.331	-2.378	-0.147	0.046	0.206	2.160
	S2	0.017	0.384	-2.334	-0.207	0.012	0.239	2.262
	S3	-0.022	0.342	-2.496	-0.194	-0.042	0.136	2.124
	S4	-0.015	0.437	-2.001	-0.258	-0.041	0.199	2.992
Model 14	P	0.016	0.285	-1.680	-0.163	0.017	0.196	1.717
	S1	0.010	0.331	-1.753	-0.163	0.022	0.197	1.747
	S2	-0.024	0.370	-1.979	-0.244	-0.031	0.199	1.799
	S3	-0.007	0.335	-1.969	-0.178	-0.024	0.152	2.052
	S4	-0.013	0.431	-2.313	-0.253	-0.037	0.203	2.241

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

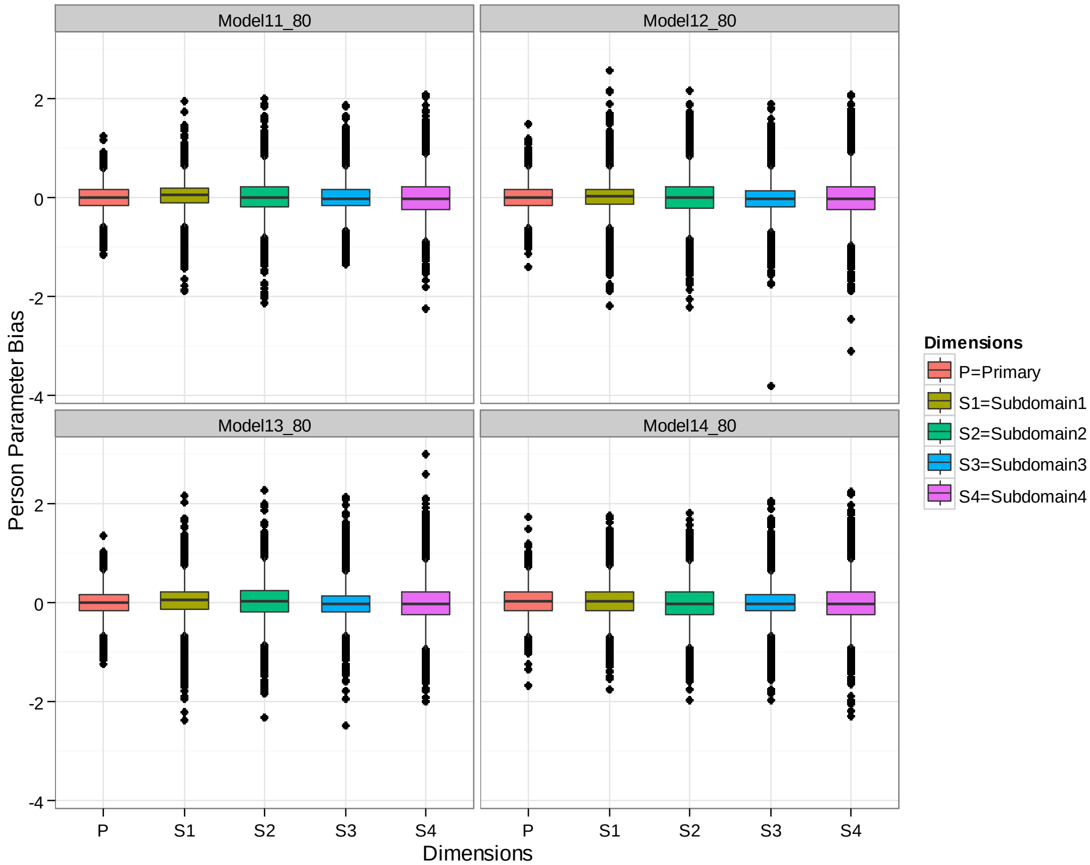

Figure 4.6: Theta Parameter Estimate Bias of 80 Items (Model 1)

4.2 Parameter Estimate Bias of Model 2

Model 1 was designed to have orthogonality violation between the third and the fourth subdomains, while Model 2 was designed to have orthogonality violation among all subdomains. Levels of the orthogonality violation were also labeled from the least to most severe violations. The specific model of Model 21 had the smallest correlation violation ($r \leq 0.1$), while Model 24 had the largest violation ($r \geq 0.5$). Results are presented in terms of item parameter and person parameter across 40-, 60-, and 80- item.

4.2.1 Item Parameter Estimate Bias of 40 Items

As with Model 1, to examine the parameter estimate recovery, the bias was computed by subtracting the true parameters from the estimated parameters. The bias scores were then summarized by the mean, the standard deviation, and the quartiles of the bias scores. Table 4.7 reports the discrimination and intercept parameter estimate bias of 40 items. As shown in the table, the intercept parameters could be accurately recovered because the mean bias scores across all specific models were all centered around zero ($\overline{X}_{Model21} = 0.002$, $\overline{X}_{Model22} = 0.009$, $\overline{X}_{Model23} = 0.057$, $\overline{X}_{Model24} = 0.014$), and the variability of the bias scores was also small ($\sigma_{Model21} = 0.047$, $\sigma_{Model22} = 0.036$, $\sigma_{Model23} = 0.042$, $\sigma_{Model24} = 0.053$). The results demonstrated that levels of orthogonality had no effect on the intercept parameter estimates.

The mean bias between the estimated and the true parameters became larger as the orthogonality violation became larger (e.g., $\overline{X}_{P\&Model21} = 0.027$, $\overline{X}_{P\&Model22} = 0.108$, $\overline{X}_{P\&Model23} = 0.178$, $\overline{X}_{P\&Model24} = 0.299$; $\overline{X}_{S1\&Model21} = -0.029$, $\overline{X}_{S1\&Model22} = -0.076$, $\overline{X}_{S1\&Model23} = -0.204$, $\overline{X}_{S1\&Model24} = -0.372$) . In other words, the mean bias of Model 22 was larger than that of Model 21, while the mean bias scores of Model 23 and Model 24 were larger than any other models with smaller orthogonality violations. In addition, the variability of the bias scores also became larger as the orthogonality violations became more severe (e.g., $\sigma_{P\&Model21} = 0.032$, $\sigma_{P\&Model22} = 0.085$, $\sigma_{P\&Model23} = 0.156$, $\sigma_{P\&Model24} = 0.244$. These results indicated that the orthogonality violation had an effect on discrimination estimate accuracy.

Figure 4.7 is the graphical display of the discrimination and intercept parameters for 40 items. As shown by the figure, discrimination parameter estimates of all subdomains were underestimated because all of the bias scores between the estimated and the true scores were below zero.

Table 4.7: Discrimination and Intercept Parameter Estimate Bias of 40 Items (Model 2)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 21	Int	0.002	0.047	-0.230	-0.004	0.001	0.014	0.161
	P	0.027	0.032	-0.057	0.005	0.031	0.051	0.075
	S1	-0.029	0.022	-0.065	-0.040	-0.026	-0.012	0.001
	S2	-0.058	0.020	-0.092	-0.068	-0.059	-0.052	-0.022
	S3	-0.057	0.022	-0.089	-0.071	-0.054	-0.045	-0.023
	S4	-0.091	0.076	-0.233	-0.116	-0.060	-0.042	-0.021
Model 22	Int	0.009	0.036	-0.138	0.001	0.008	0.017	0.140
	P	0.108	0.085	-0.080	0.038	0.123	0.181	0.265
	S1	-0.076	0.028	-0.119	-0.100	-0.072	-0.062	-0.031
	S2	-0.157	0.049	-0.237	-0.190	-0.159	-0.139	-0.075
	S3	-0.125	0.041	-0.194	-0.147	-0.129	-0.117	-0.058
	S4	-0.149	0.074	-0.275	-0.198	-0.125	-0.093	-0.050
Model 23	Int	0.057	0.042	-0.029	0.034	0.053	0.075	0.224
	P	0.178	0.156	-0.158	0.057	0.176	0.325	0.392
	S1	-0.204	0.077	-0.322	-0.270	-0.204	-0.135	-0.113
	S2	-0.272	0.090	-0.411	-0.343	-0.272	-0.222	-0.122
	S3	-0.224	0.069	-0.313	-0.265	-0.243	-0.203	-0.106
	S4	-0.281	0.108	-0.431	-0.369	-0.275	-0.205	-0.097
Model 24	Int	0.014	0.053	-0.142	-0.005	0.011	0.045	0.125
	P	0.299	0.244	-0.194	0.104	0.303	0.518	0.685
	S1	-0.372	0.118	-0.577	-0.454	-0.361	-0.275	-0.215
	S2	-0.425	0.143	-0.669	-0.523	-0.437	-0.339	-0.199
	S3	-0.362	0.124	-0.533	-0.436	-0.373	-0.316	-0.153
	S4	-0.467	0.175	-0.674	-0.634	-0.434	-0.367	-0.168

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

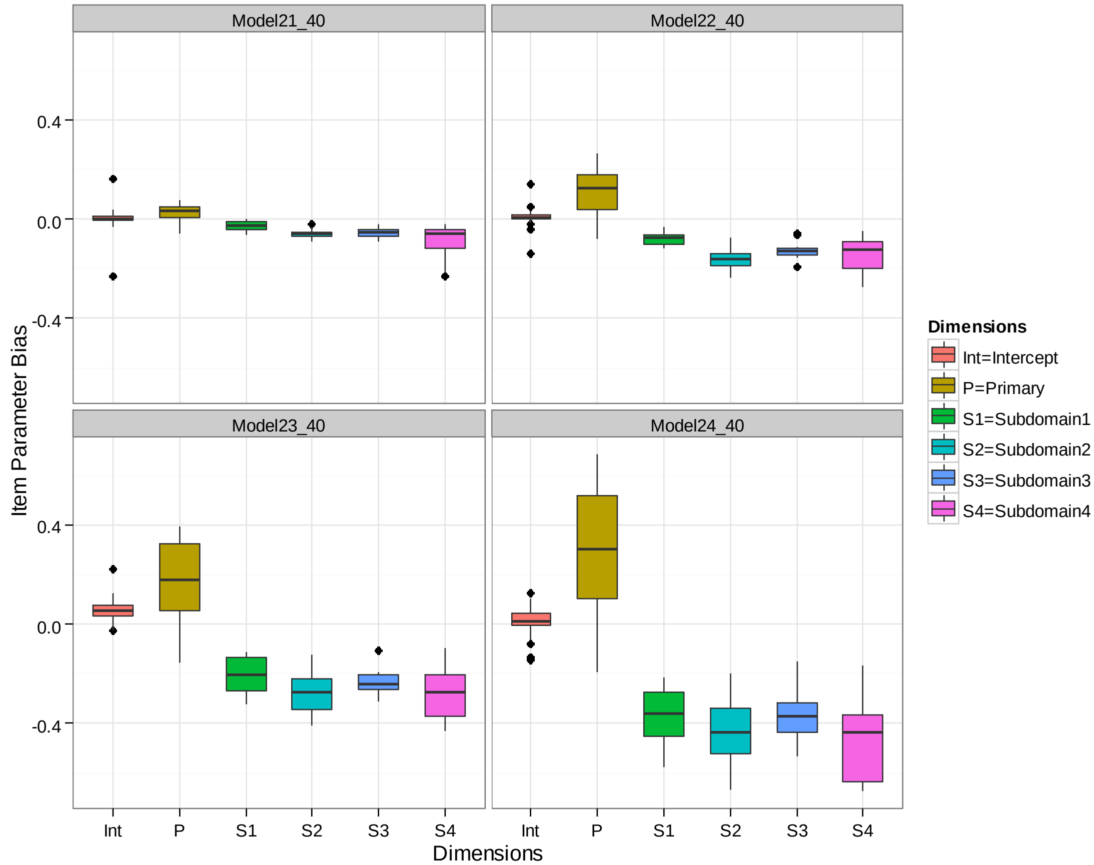

Figure 4.7: Discrimination and Intercept Parameter Estimate Bias of 40 Items (Model 2)

4.2.2 Item Parameter Estimate Bias of 60 Items

Table 4.8 illustrates the discrimination and intercept parameter estimate bias of 60 items. The results indicated the same trend as those of 40 items. The intercept parameters could be accurately recovered because the mean bias was centered around zero and the variance was small ($\overline{X}_{Model21} = 0.015$, $\overline{X}_{Model22} = 0.038$, $\overline{X}_{Model23} = 0.012$, $\overline{X}_{Model24} = -0.005$; $\sigma_{Model21} = 0.063$, $\sigma_{Model22} = 0.050$, $\sigma_{Model23} = 0.041$, $\sigma_{Model24} = 0.038$). Levels of orthogonality violations had no effect on the intercept parameter estimates.

The mean bias and variance of the discrimination parameters became larger as the orthogonality violations became larger, which was consistent with those of 40 items ($\overline{X}_{P\&Model21} = 0.031$, $\overline{X}_{P\&Model22} = 0.110$, $\overline{X}_{P\&Model23} = 0.213$, $\overline{X}_{P\&Model24} = 0.315$; $\sigma_{P\&Model21} = 0.040$,

$\sigma_{P\&Model22} = 0.064$, $\sigma_{P\&Model23} = 0.148$, $\sigma_{P\&Model24} = 0.235$. Figure 4.8 is the graphical display of the discrimination and intercept parameters for 60 items. As with the results of 40 items, the discrimination parameters of all subdomains were underestimated because all the bias scores were below zero. There were not noticeable differences in parameter estimates between 40 and 60 items.

Table 4.8: Discrimination and Intercept Parameter Estimate Bias of 60 Items (Model 2)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 21	Int	0.015	0.063	-0.069	-0.015	0.006	0.030	0.382
	P	0.031	0.040	-0.068	-0.001	0.037	0.064	0.094
	S1	-0.022	0.029	-0.073	-0.044	-0.016	0.004	0.008
	S2	-0.054	0.034	-0.139	-0.070	-0.052	-0.025	-0.009
	S3	-0.034	0.017	-0.076	-0.040	-0.033	-0.022	-0.005
	S4	-0.097	0.074	-0.323	-0.105	-0.090	-0.050	-0.030
Model 22	Int	0.038	0.050	-0.020	0.017	0.023	0.045	0.331
	P	0.110	0.064	-0.031	0.063	0.119	0.165	0.212
	S1	-0.063	0.030	-0.120	-0.082	-0.051	-0.039	-0.027
	S2	-0.159	0.069	-0.281	-0.193	-0.165	-0.104	-0.056
	S3	-0.122	0.053	-0.215	-0.167	-0.112	-0.085	-0.043
	S4	-0.147	0.078	-0.343	-0.181	-0.133	-0.092	-0.051
Model 23	Int	0.012	0.041	-0.056	-0.011	0.004	0.017	0.160
	P	0.213	0.148	-0.157	0.102	0.211	0.338	0.458
	S1	-0.181	0.073	-0.351	-0.236	-0.153	-0.122	-0.091
	S2	-0.313	0.131	-0.542	-0.394	-0.312	-0.212	-0.109
	S3	-0.267	0.100	-0.433	-0.340	-0.255	-0.216	-0.095
	S4	-0.316	0.136	-0.645	-0.391	-0.316	-0.224	-0.139
Model 24	Int	-0.005	0.038	-0.101	-0.019	-0.006	0.003	0.153
	P	0.315	0.235	-0.252	0.140	0.317	0.521	0.681
	S1	-0.351	0.132	-0.614	-0.461	-0.335	-0.239	-0.195
	S2	-0.479	0.196	-0.812	-0.591	-0.495	-0.330	-0.172
	S3	-0.426	0.158	-0.680	-0.547	-0.396	-0.345	-0.163
	S4	-0.472	0.192	-0.902	-0.594	-0.489	-0.328	-0.216

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

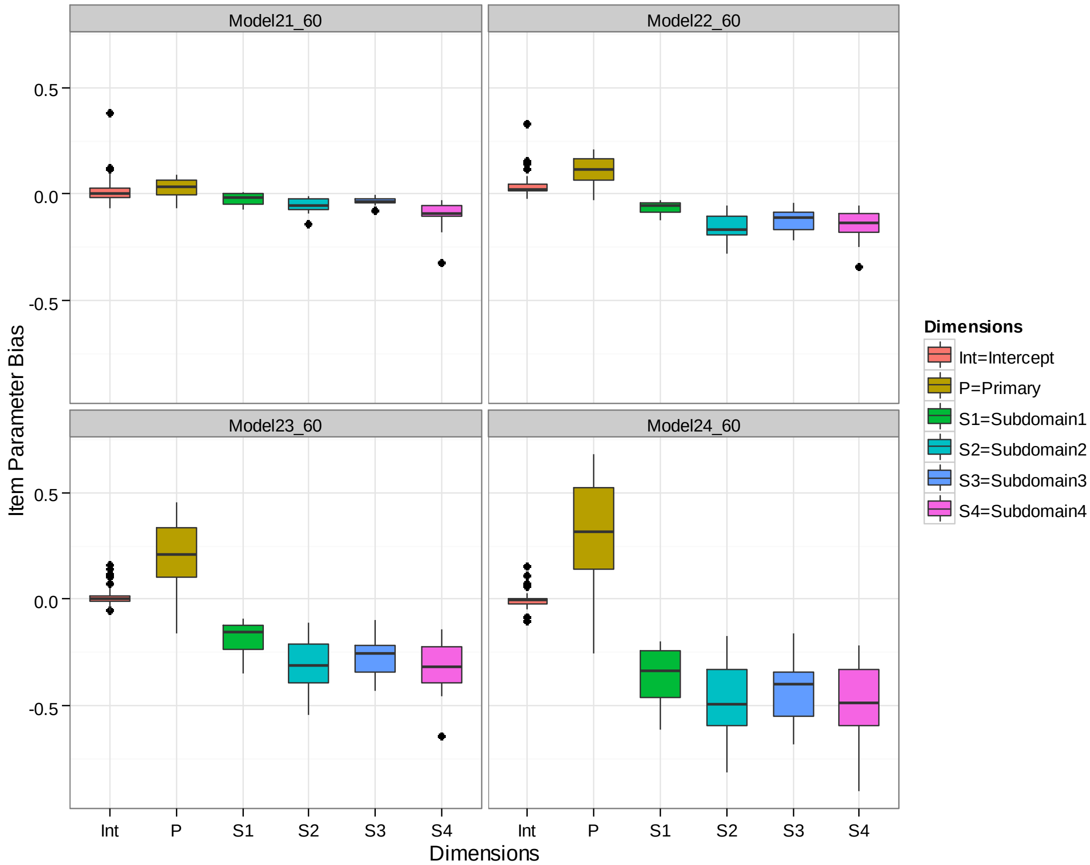

Figure 4.8: Discrimination and Intercept Parameter Estimate Bias of 60 Items (Model 2)

4.2.3 Item Parameter Estimate Bias of 80 Items

Table 4.9 reports the intercept and discrimination parameter bias of 80 items. The results followed the same trend as those with 40 and 60 items. Specific models had little influence on intercept parameter estimates because the mean bias scores were around zero and the variance of the bias scores was small ($\overline{X}_{Model21} = -0.011$, $\overline{X}_{Model22} = 0.021$, $\overline{X}_{Model23} = 0.006$, $\overline{X}_{Model24} = -0.015$; $\sigma_{Model21} = 0.036$, $\sigma_{Model22} = 0.027$, $\sigma_{Model23} = 0.043$, $\sigma_{Model24} = 0.030$). Intercept parameters could be accurately recovered.

The bias of the discrimination parameters increased as the orthogonality violations became severe ($\overline{X}_{P\&Model21} = 0.022$, $\overline{X}_{P\&Model22} = 0.067$, $\overline{X}_{P\&Model23} = 0.149$, $\overline{X}_{P\&Model24} = 0.240$; $\sigma_{P\&Model21} = 0.027$, $\sigma_{P\&Model22} = 0.038$, $\sigma_{P\&Model23} = 0.095$, $\sigma_{P\&Model24} = 0.169$). Model 21 had the smallest bias, while Model 24 had the largest bias. Figure 4.9 is the graphical display of discrimination and intercept parameters for 80 items. As with the discrimination estimates of 40 and 60 items, parameter estimates of all subdomains were underestimated because all the bias scores were below zero.

Compared to the bias range of 40 and 60 items, the bias range of 80 items was smaller (e.g., $Q1_{S4\&40} = -0.634$, $Q3_{S4\&40} = -0.518$; $Q1_{S4\&60} = -0.594$, $Q3_{S4\&60} = -0.521$; $Q1_{S4\&80} = -0.497$, $Q3_{S4\&80} = -0.386$). The number of items may have an effect on the parameter estimate accuracy. Further analysis is needed to examine whether the effects are significant.

Table 4.9: Discrimination and Intercept Parameter Discovery of 80 Items (Model 2)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 21	Int	-0.011	0.036	-0.131	-0.027	-0.010	0.006	0.119
	P	0.022	0.027	-0.043	0.005	0.017	0.033	0.091
	S1	-0.011	0.020	-0.068	-0.017	-0.005	0.004	0.006
	S2	-0.067	0.043	-0.149	-0.095	-0.048	-0.034	-0.019
	S3	-0.042	0.018	-0.089	-0.051	-0.041	-0.029	-0.012
	S4	-0.066	0.029	-0.109	-0.084	-0.061	-0.043	-0.020
Model 22	Int	0.021	0.027	-0.066	0.007	0.018	0.036	0.096
	P	0.067	0.038	0.001	0.037	0.062	0.083	0.171
	S1	-0.075	0.034	-0.145	-0.099	-0.080	-0.053	-0.014
	S2	-0.134	0.047	-0.224	-0.158	-0.128	-0.106	-0.049
	S3	-0.061	0.039	-0.134	-0.091	-0.057	-0.031	-0.004
	S4	-0.107	0.040	-0.176	-0.132	-0.104	-0.078	-0.028
Model 23	Int	0.006	0.043	-0.097	-0.012	-0.001	0.021	0.178
	P	0.149	0.095	-0.056	0.053	0.175	0.229	0.292
	S1	-0.170	0.073	-0.341	-0.215	-0.161	-0.111	-0.057
	S2	-0.230	0.076	-0.385	-0.275	-0.219	-0.180	-0.113
	S3	-0.168	0.070	-0.292	-0.214	-0.169	-0.111	-0.038
	S4	-0.247	0.083	-0.373	-0.304	-0.241	-0.185	-0.072
Model 24	Int	-0.015	0.030	-0.075	-0.037	-0.012	0.001	0.093
	P	0.240	0.169	-0.080	0.072	0.275	0.386	0.565
	S1	-0.326	0.138	-0.584	-0.438	-0.302	-0.215	-0.110
	S2	-0.378	0.137	-0.642	-0.481	-0.396	-0.248	-0.197
	S3	-0.343	0.131	-0.570	-0.414	-0.351	-0.234	-0.102
	S4	-0.392	0.138	-0.623	-0.497	-0.369	-0.274	-0.121

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile, Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

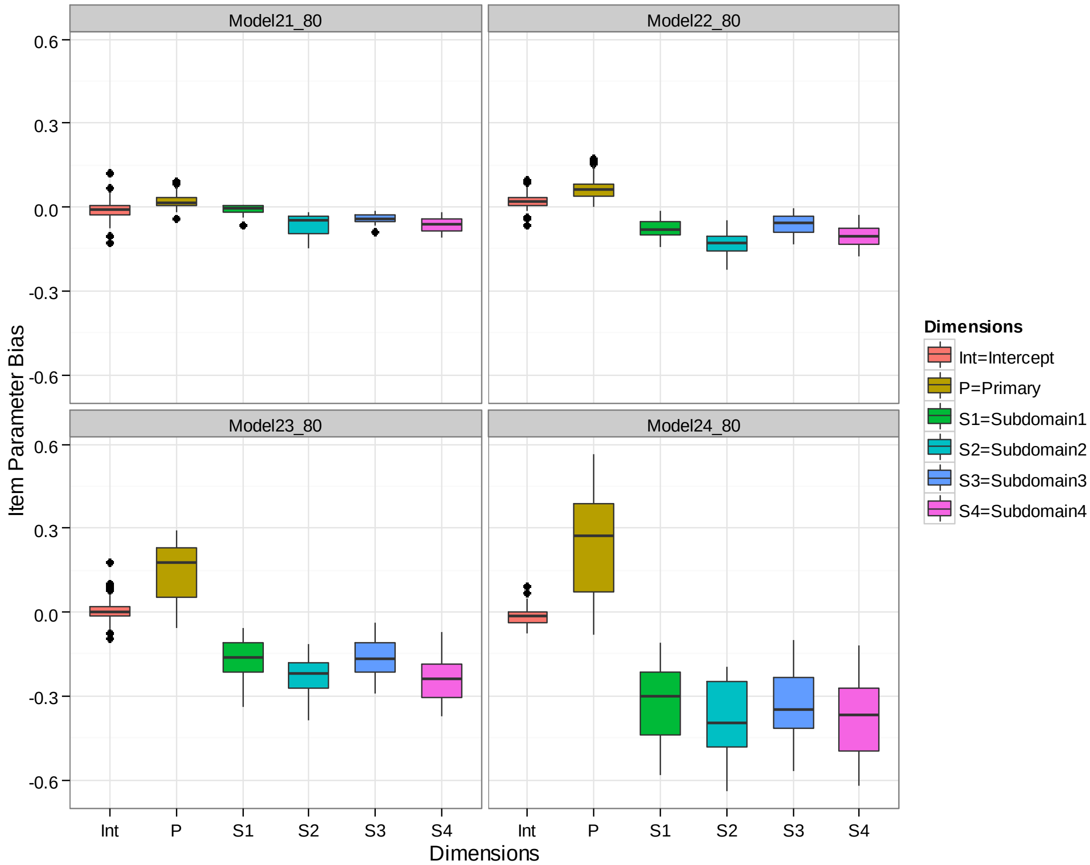

Figure 4.9: Discrimination and Intercept Parameter Estimate Bias of 80 Items (Model 2)

4.2.4 Person Parameter Estimate Bias of 40 Items

Table 4.10 demonstrates the theta parameter estimate bias of 40 items when the orthogonality violations exist among all subdomains. As with the theta results in Model 1 (i.e., the orthogonality violations exist only between the third and fourth subdomains), the estimated theta parameters were much different from the true parameters because the range of bias scores was large ($min = -3.581$and $max = -3.628$). There are many outliers in the bias scores (see Figure 4.10). Specific models did not affect the theta parameters estimates.

The mean bias of all dimensions across specific models did not fluctuate much. All were centered at zero (e.g., $\overline{X}_{P\&Model21} = 0.004$, $\overline{X}_{S1\&Model21} = 0.009$; $\overline{X}_{S2\&Model21} = 0.001$, $\overline{X}_{S3\&Model21} = -0.007$; $\overline{X}_{S4\&Model21} = -0.006$). The standard deviation of the primary dimension was the smallest among all dimensions, making the parameter estimates of the primary dimension the most trustworthy ($\sigma_{P\&Model21} = 0.357$, $\sigma_{S1\&Model21} = 0.425$; $\sigma_{S2\&Model21} = 0.458$, $\sigma_{S3\&Model21} = 0.602$, $\sigma_{S4\&Model21} = 0.493$). The third dimension exhibited the most bias from the true parameters, which was consistent with the 40-item estimates of Model 1.

Figure 4.10 is the graphical display of the theta parameter bias of 40 items. As shown by the figure, the bias scores had a wide range. Because the bias scores of the primary dimension were smaller than those of other dimensions, the estimates of the the primary dimension were most trustworthy. In addition, the distributions of the bias cores were approximately symmetric with a mean bias scores around zero.

Table 4.10: Theta Parameter Estimate Bias of 40 Items (Model 2)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 21	P	0.004	0.357	-1.558	-0.231	-0.005	0.233	1.592
	S1	0.009	0.425	-2.439	-0.198	0.038	0.257	2.141
	S2	0.001	0.458	-2.704	-0.230	0.052	0.294	2.103
	S3	-0.007	0.602	-2.524	-0.378	0.032	0.409	2.516
	S4	-0.006	0.493	-1.827	-0.325	-0.004	0.310	2.093
Model 22	P	0.003	0.400	-1.658	-0.263	0.001	0.258	1.573
	S1	0.004	0.450	-2.625	-0.214	0.046	0.267	2.087
	S2	-0.004	0.493	-2.898	-0.238	0.075	0.306	1.659
	S3	-0.002	0.634	-2.950	-0.391	0.038	0.424	2.798
	S4	0.015	0.515	-2.110	-0.319	0.003	0.348	2.185
Model 23	P	0.040	0.488	-1.807	-0.283	0.049	0.363	2.096
	S1	0.019	0.493	-2.461	-0.236	0.045	0.308	2.628
	S2	0.009	0.536	-3.547	-0.250	0.072	0.342	2.157
	S3	0.009	0.693	-3.169	-0.411	0.030	0.473	3.119
	S4	0.030	0.571	-1.905	-0.353	0.013	0.405	2.386
Model 24	P	0.008	0.580	-1.911	-0.393	0.005	0.395	2.079
	S1	0.007	0.564	-3.581	-0.298	0.020	0.328	2.665
	S2	0.021	0.620	-2.702	-0.299	0.082	0.412	2.460
	S3	0.013	0.792	-3.187	-0.484	0.057	0.531	3.259
	S4	-0.001	0.679	-2.520	-0.458	-0.008	0.441	3.628

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile, Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

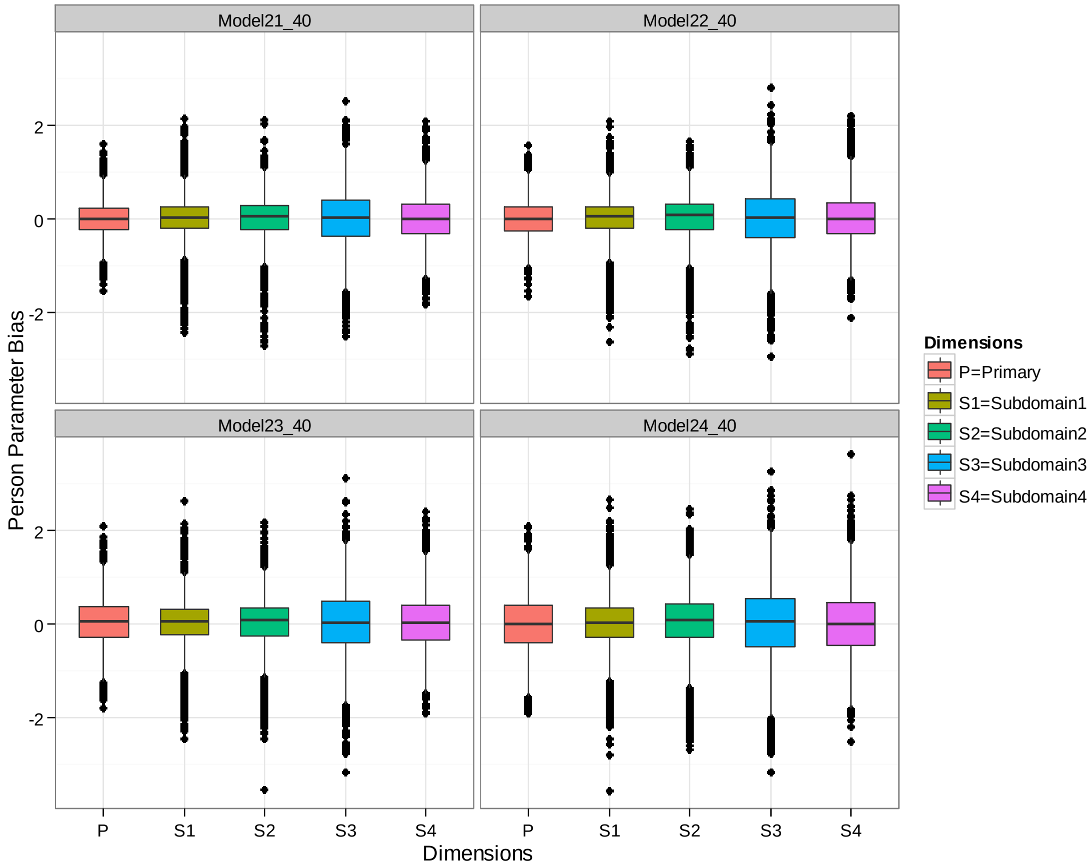

Figure 4.10: Theta Parameter Estimate Bias of 40 Items (Model 2)

4.2.5 Person Parameter Estimate Bias of 60 Items

Table 4.11 demonstrates the theta parameter estimate bias of 60 items. The results followed exactly the same pattern as the model with 40 items. The bias scores had a wide range between -3.789 and 3.267. The distribution of the bias scores was approximately symmetric, with means all centered at zero. The bias scores and variance of the primary dimensions were the smallest relative to other dimensions; thus, the parameter estimates of the primary dimensions were most trustworthy. Figure 4.11 is the graphical display of the theta parameters of 60 items.

Table 4.11: Theta Parameter Estimate Bias of 60 Items (Model 2)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 21	P	0.007	0.357	-1.561	-0.227	0.013	0.239	1.594
	S1	0.024	0.397	-2.214	-0.189	0.052	0.266	2.164
	S2	-0.029	0.430	-2.536	-0.281	-0.030	0.219	2.483
	S3	-0.023	0.402	-2.059	-0.254	-0.022	0.210	2.130
	S4	0.014	0.489	-2.024	-0.289	0.013	0.305	2.743
Model 22	P	0.021	0.412	-2.037	-0.256	0.022	0.292	1.840
	S1	0.011	0.436	-2.954	-0.224	0.040	0.273	2.344
	S2	0.004	0.460	-2.182	-0.266	0.006	0.284	2.352
	S3	-0.005	0.424	-3.109	-0.245	0.003	0.244	1.994
	S4	0.013	0.539	-3.152	-0.306	-0.002	0.330	2.415
Model 23	P	-0.004	0.507	-1.886	-0.332	-0.004	0.333	1.717
	S1	0.007	0.488	-2.579	-0.268	0.047	0.316	2.231
	S2	0.014	0.539	-2.910	-0.302	0.012	0.329	2.830
	S3	-0.008	0.505	-2.163	-0.286	-0.005	0.279	2.735
	S4	0.008	0.614	-3.198	-0.366	-0.009	0.365	2.887
Model 24	P	-0.008	0.593	-2.315	-0.407	-0.001	0.383	1.819
	S1	-0.005	0.566	-2.948	-0.348	0.028	0.359	3.021
	S2	-0.018	0.645	-3.342	-0.406	-0.013	0.375	3.175
	S3	0.001	0.609	-2.971	-0.358	0.003	0.362	3.267
	S4	0.006	0.725	-3.789	-0.421	-0.003	0.442	3.118

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

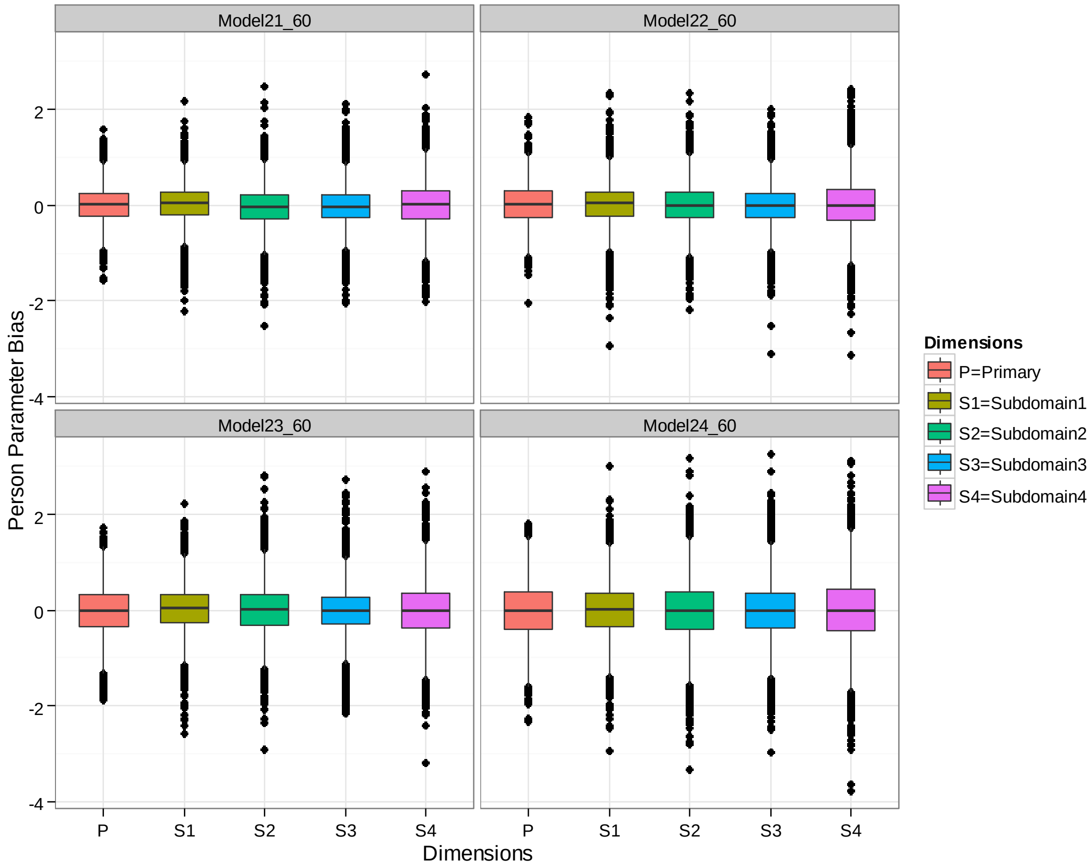

Figure 4.11: Theta Parameter Estimate Bias of 60 Items (Model 2)

4.2.6 Person Parameter Estimate Bias of 80 Items

Table 4.12 describes the theta parameter bias of 80 items. The results shared much similarity with the models of 40 and 60 items. The mean bias scores of all dimensions and across all the models were nearly identical. The bias range of the primary dimension was smaller compared with the other subdomains. Figure 4.12 is the graphical display of the person parameter bias of 80 items.

As with the results of Model 1, the bias range of 80 items was smaller than the ones with 40 and 60 items ($Q1_{40} = 0.484$, $Q3_{40} = 0.531$; $Q1_{60} = 0.421$, $Q3_{60} = 0.442$; $Q1_{80} = 0.349$, $Q3_{80} = 0.259$). These results indicated that the number of items may influence the theta parameter estimate accuracy; however, the specific models did not affect the theta estimates.

Table 4.12: Theta Parameter Estimate Bias of 80 Items (Model 2)

		$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Model 21	P	-0.009	0.256	-1.478	-0.162	-0.004	0.146	1.272
	S1	0.018	0.330	-2.199	-0.134	0.040	0.197	2.008
	S2	-0.023	0.374	-1.905	-0.232	-0.035	0.181	2.151
	S3	-0.012	0.317	-1.804	-0.187	-0.026	0.149	2.015
	S4	0.002	0.414	-1.843	-0.238	-0.023	0.221	2.156
Model 22	P	0.012	0.291	-1.483	-0.163	0.014	0.197	1.743
	S1	0.008	0.320	-1.847	-0.145	0.033	0.182	2.019
	S2	-0.004	0.384	-2.227	-0.209	-0.014	0.207	2.969
	S3	0.007	0.341	-1.966	-0.182	-0.006	0.177	2.000
	S4	-0.001	0.432	-2.710	-0.253	-0.028	0.227	3.079
Model 23	P	0.001	0.373	-1.817	-0.243	0.004	0.238	1.541
	S1	0.034	0.368	-2.753	-0.145	0.061	0.241	1.864
	S2	-0.008	0.431	-2.603	-0.251	-0.022	0.235	2.663
	S3	-0.001	0.393	-2.116	-0.220	-0.025	0.202	2.126
	S4	-0.015	0.493	-3.436	-0.292	-0.037	0.232	2.857
Model 24	P	0.010	0.453	-1.754	-0.283	0.012	0.315	1.498
	S1	-0.014	0.420	-2.467	-0.234	0.013	0.230	2.476
	S2	-0.022	0.493	-2.444	-0.316	-0.036	0.260	3.099
	S3	-0.031	0.419	-2.763	-0.279	-0.047	0.187	2.865
	S4	-0.023	0.542	-2.613	-0.349	-0.062	0.259	3.182

Note: Min=Minimum, Max =Maximum, Q1 = 1^{st} quartile,
Q2 = 2^{nd} quartile, Q3 = 3^{rd} quartile

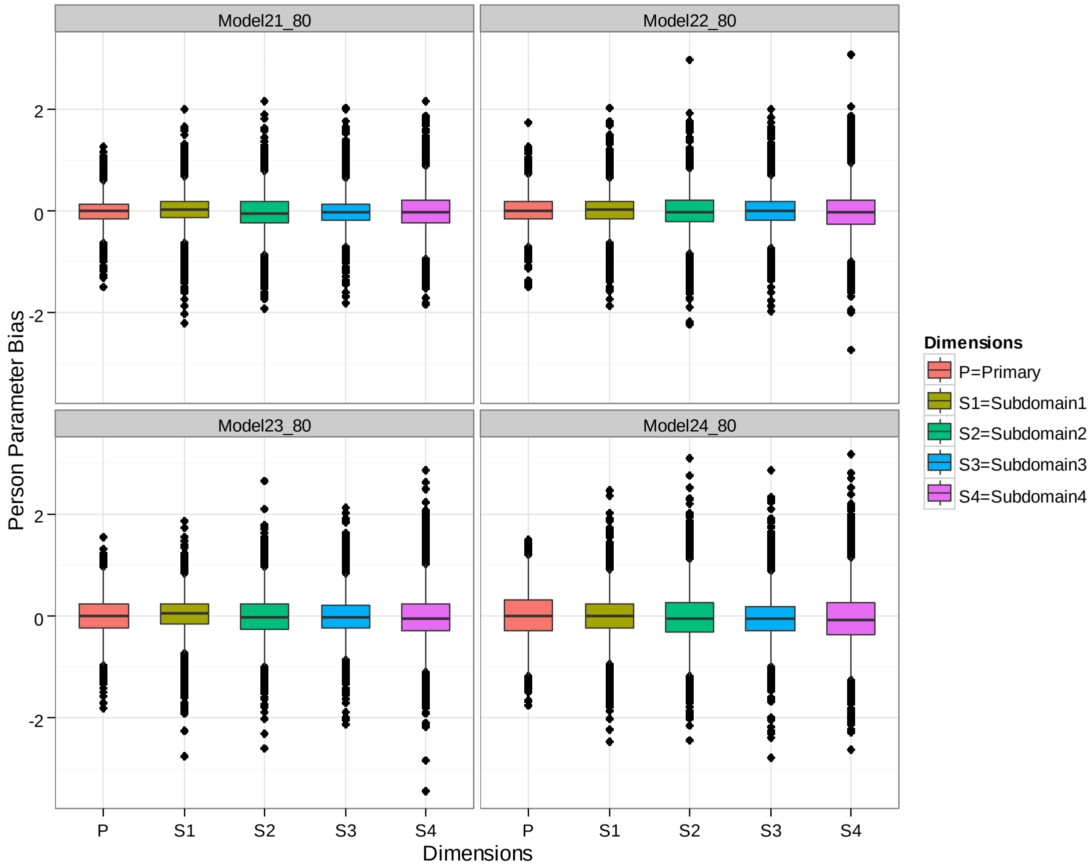

Figure 4.12: Theta Parameter Estimate Bias of 80 Items (Model 2)

4.3 Effects of Models and Numbers of Items

The previous two sections described the parameter estimate bias for all 24 cases. This current section presents the magnitude of the model and numbers of items effects on parameter estimate bias. The purpose of this section, is to test whether parameter estimates obtained from violations between two subdomains were significantly different from the parameters from violations among all subdomains. The number of Items was also tested to examine parameter estimate accuracy. The mean bias and the mean RMSE were summarized in terms of a, d, θ parameters respectively. Then ANOVA was conducted to test whether the model and the number of items influenced item and person parameter estimates.

4.3.1 Mean Bias and Mean RMSE of *a* Parameters

Table 4.13 indicates the mean bias and the mean RMSE for *a* parameters. As shown in the table, the mean bias of Model 2 was larger than that of Model 1 because every value in Model 2 was larger than the corresponding value in Model 1($\overline{X}_{Bias\&Model1}$= -0.013, $\overline{X}_{RMSE\&Model2}$= -0.028). The mean RMSE also demonstrated the same result: the mean RMSE of Model 2 was larger than that of Model 1($\overline{X}_{RMSE\&Model1}$= 0.103, $\overline{X}_{RMSE\&Model2}$= 0.199). Results indicated that parameters estimate bias of the model with violations among all subdomains was larger than the model with violations between two subdomains.

The number of items indicated a small influence on parameter estimate accuracy because few differences were found in both the mean bias and the mean RMSE across 40, 60, and 80 items ($\overline{X}_{Bias\&40}$= -0.022, $\overline{X}_{Bias\&60}$= -0.023, $\overline{X}_{Bias\&80}$= -0.016 ; $\overline{X}_{RMSE\&40}$= 0.159, $\overline{X}_{RMSE\&60}$= 0.162, $\overline{X}_{RMSE\&80}$= 0.131). Figure 4.13is the graphical display of the mean bias and mean RMSE for the *a* Parameters.

Table 4.13: Mean Bias and Mean RMSE of Models and Numbers of Items for *a* Parameters

		$\overline{X}_{Bias}$ (σ)			$\overline{X}_{RMSE}$ (σ)		
		40 Items	60 Items	80 Items	40 Items	60 Items	80 Items
Model 1	Model 11	-0.009 (0.051)	-0.022 (0.039)	0.001 (0.045)	0.080 (0.038)	0.073 (0.033)	0.072 (0.023)
	Model 12	-0.001 (0.054)	-0.012 (0.061)	0.005 (0.057)	0.084 (0.037)	0.081 (0.042)	0.080 (0.026)
	Model 13	-0.019 (0.093)	-0.018 (0.099)	-0.005 (0.086)	0.105 (0.057)	0.106 (0.059)	0.097 (0.042)
	Model 14	-0.036 (0.174)	-0.025 (0.194)	-0.016 (0.117)	0.164 (0.097)	0.172 (0.114)	0.120 (0.061)
Model 2	Model 21	-0.016 (0.058)	-0.011 (0.062)	-0.012 (0.047)	0.084 (0.039)	0.082 (0.039)	0.073 (0.027)
	Model 22	-0.010 (0.138)	-0.006 (0.135)	-0.014 (0.092)	0.144 (0.061)	0.136 (0.063)	0.104 (0.041)
	Model 23	-0.034 (0.248)	-0.028 (0.277)	-0.028 (0.198)	0.236 (0.108)	0.256 (0.128)	0.191 (0.084)
	Model 24	-0.054 (0.407)	-0.059 (0.428)	-0.06 (0.337)	0.375 (0.177)	0.390 (0.196)	0.312 (0.152)

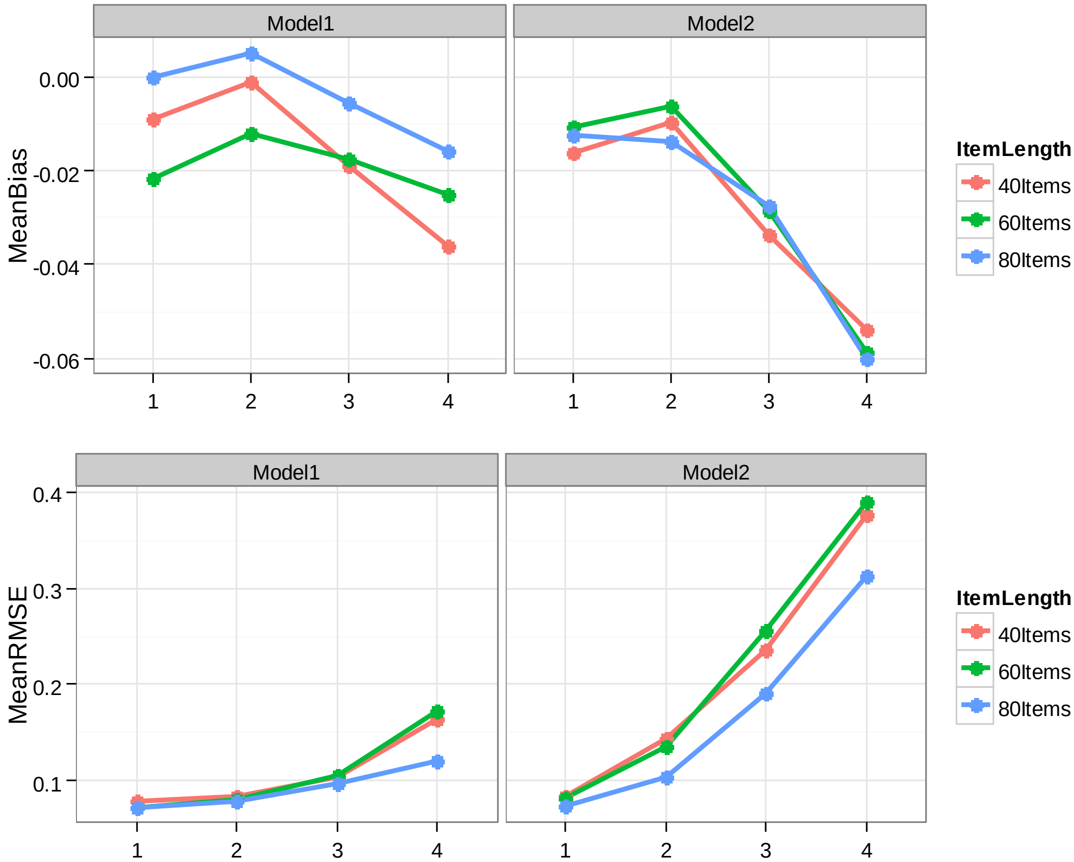

Figure 4.13: Mean Bias and Mean RMSE of Models and Numbers of Items for *a* Parameters

4.3.2 Mean Bias and Mean RMSE of *d* parameters

Table 4.14 reports the mean bias and mean RMSE of the *d* parameters. The mean bias of Model 1 was similar to that of Model 2 ($\overline{X}_{Bias\&Model1}$= -0.007, $\overline{X}_{Bias\&Model2}$= -0.012). The mean RMSE of the Model 1 also indicated little difference from Model 2 ($\overline{X}_{RMSE\&Model1}$= -0.065, $\overline{X}_{RMSE\&Model2}$= -0.067). The results revealed that the intercept parameter estimates may not significantly differ from the true parameters regardless of the violations between two subdomains or among all dimensions.

The number of items also did not have a notable effect on *d* parameter estimates. The mean bias scores were not much different from each other across the different numbers of items ($\overline{X}_{Bias\&40}$= -0.012, $\overline{X}_{Bias\&60}$= -0.017, $\overline{X}_{Bias\&80}$= -0.001), and the mean RMSE scores were also not significantly different from each other ($\overline{X}_{RMSE\&40}$= 0.068, $\overline{X}_{RMSE\&60}$= 0.067, $\overline{X}_{RMSE\&80}$= 0.063). Figure 4.14 is the graphical display of the mean bias and mean RMSE for the *d* parameters.

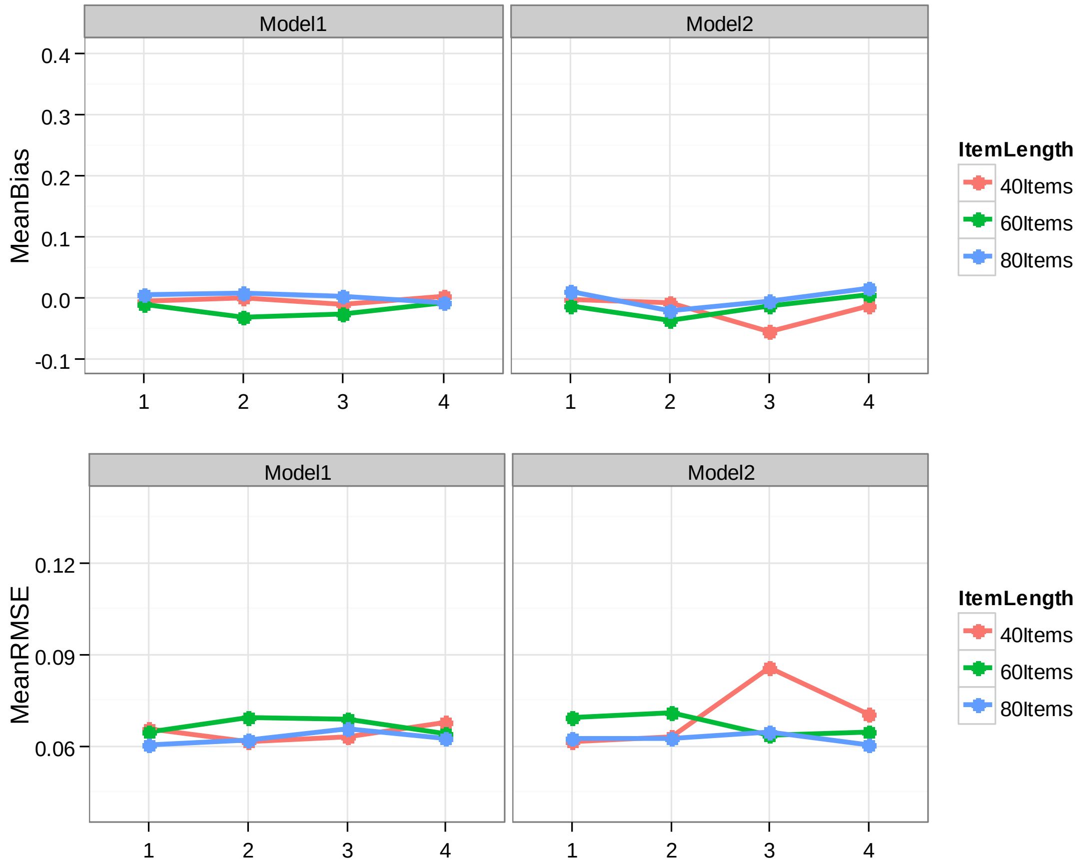

Figure 4.14: Mean Bias and Mean RMSE of Models and Numbers of Items for *d* Parameters

4.3.3 Mean Bias and Mean RMSE of θ Parameters

Table 4.15 presents the mean bias and the mean RMSE of models and numbers of items for the θ parameters. The mean bias scores of Model 1 and Model 2 had small differences in the θ parameter estimate ($\overline{X}_{Bias\&Model1}$= -0.001, $\overline{X}_{Bias\&Model2}$= -0.002), and the mean RMSE of Model 1 and Model 2 also indicated small differences in the theta estimates ($\overline{X}_{RMSE\&Model1}$= 0.565, $\overline{X}_{RMSE\&Model2}$= 0.605). These outcomes suggest that the model with violations among all subdomains might not be significant different from the model with violations between two subdomains in parameter estimates.

Table 4.14: Mean Bias and Mean RMSE of Models and Numbers of Items for d Parameters

		$\overline{X}_{Bias}$ (σ)			$\overline{X}_{RMSE}$ (σ)		
		40 Items	60 Items	80 Items	40 Items	60 Items	80 Items
Model 1	Model 11	-0.005 (0.049)	-0.011 (0.048)	0.005 (0.031)	0.066 (0.052)	0.065 (0.053)	0.061 (0.026)
	Model 12	0.001 (0.032)	-0.031 (0.053)	0.008 (0.036)	0.062 (0.050)	0.070 (0.058)	0.062 (0.028)
	Model 13	-0.010 (0.032)	-0.028 (0.047)	0.003 (0.044)	0.063 (0.045)	0.069 (0.057)	0.065 (0.031)
	Model 14	0.003 (0.052)	-0.009 (0.043)	-0.008 (0.033)	0.068 (0.051)	0.064 (0.049)	0.063 (0.030)
Model 2	Model 21	-0.002 (0.047)	-0.015 (0.063)	0.011 (0.036)	0.062 (0.051)	0.069 (0.061)	0.063 (0.028)
	Model 22	-0.009 (0.036)	-0.038 (0.050)	-0.021 (0.027)	0.063 (0.047)	0.071 (0.058)	0.063 (0.027)
	Model 23	-0.057 (0.042)	-0.012 (0.041)	-0.006 (0.043)	0.086 (0.049)	0.063 (0.050)	0.064 (0.031)
	Model 24	-0.014 (0.053)	0.005 (0.038)	0.015 (0.030)	0.071 (0.047)	0.065 (0.048)	0.060 (0.025)

As for the number of items, there were not substantial differences in the mean bias of the theta parameter estimates ($\overline{X}_{Bias\&40}$= 0.005, $\overline{X}_{Bias\&60}$= 0.003, $\overline{X}_{Bias\&80}$= -0.003), and there were not much difference in the mean RMSE either ($\overline{X}_{RMSE\&40}$= 0.633, $\overline{X}_{RMSE\&60}$= 0.596, $\overline{X}_{RMSE\&80}$= 0.526). The number of items had little effect on parameter estimates. Figure 4.15 is the graphical display of the mean bias and mean RMSE for the θ parameters.

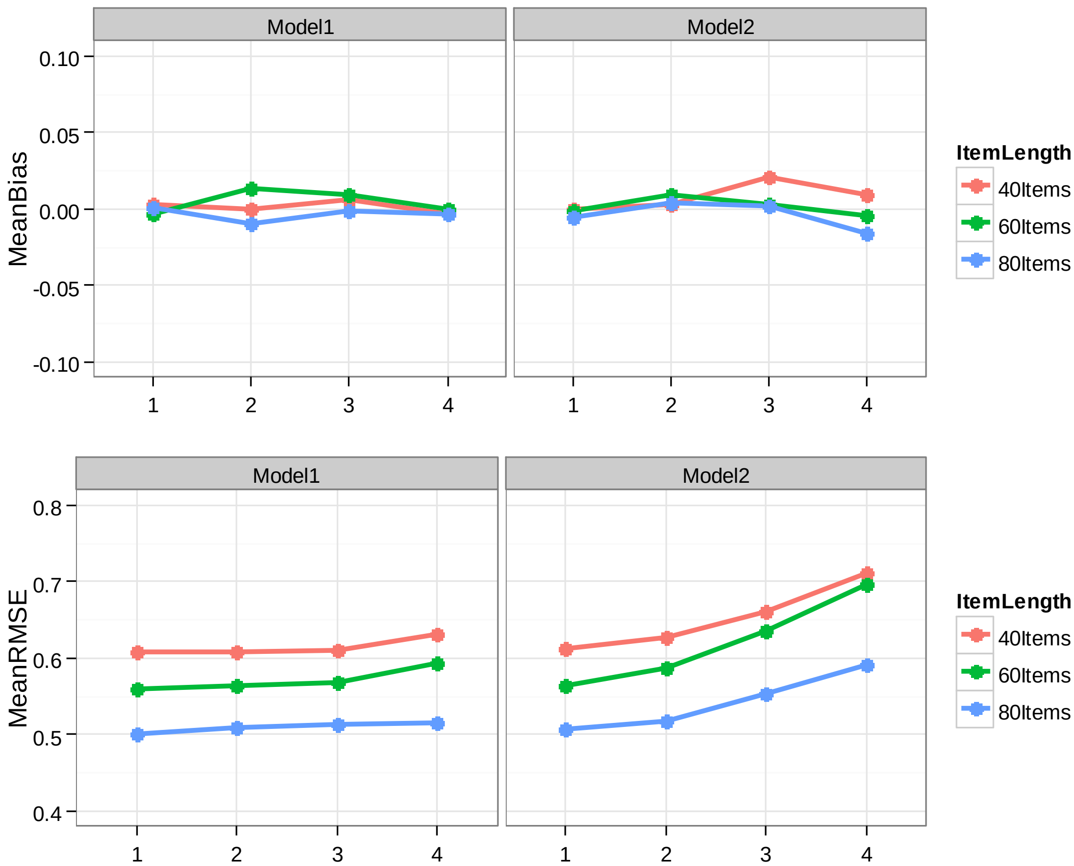

Figure 4.15: Mean Bias and Mean RMSE of Models and Items Lengths for θ Parameters

4.3.4 Models and Numbers of Items Effects on a, d, and θ Parameters

Table 4.16 reports the statistical test of models and numbers of items on the mean RMSE of the item and person parameters. As shown in the table, there was a significant effect for models, $F(1, 2874) = 470.468$, $p = 0.001$, $\eta^2 = 0.141$, a significant effect for the number of items, $F(2, 2874) = 23.179$, $p = 0.001$, $\eta^2 = 0.016$, and a significant interaction between the models and the number

Table 4.15: Mean Bias and Mean RMSE of Models and Numbers of Items for θ Parameters

		$\overline{X}_{Bias}$ (σ)			$\overline{X}_{RMSE}$ (σ)		
		40 Items	60 Items	80 Items	40 Items	60 Items	80 Items
Model 1	Model 11	0.003 (0.467)	-0.003 (0.407)	0.001 (0.332)	0.607 (0.222)	0.559 (0.192)	0.501 (0.169)
	Model 12	0.001 (0.472)	0.014 (0.414)	-0.009 (0.350)	0.608 (0.226)	0.563 (0.196)	0.509 (0.185)
	Model 13	0.007 (0.474)	0.009 (0.422)	-0.001 (0.357)	0.609 (0.225)	0.568 (0.202)	0.514 (0.186)
	Model 14	-0.004 (0.516)	-0.001 (0.470)	-0.004 (0.354)	0.631 (0.251)	0.593 (0.230)	0.516 (0.179)
Model 2	Model 21	0.001 (0.474)	-0.001 (0.418)	-0.005 (0.343)	0.612 (0.225)	0.565 (0.198)	0.507 (0.176)
	Model 22	0.003 (0.504)	0.009 (0.457)	0.005 (0.357)	0.627 (0.243)	0.588 (0.221)	0.517 (0.183)
	Model 23	0.021 (0.561)	0.004 (0.533)	0.002 (0.415)	0.660 (0.274)	0.636 (0.264)	0.553 (0.212)
	Model 24	0.010 (0.652)	-0.005 (0.630)	-0.016 (0.468)	0.711 (0.329)	0.697 (0.320)	0.590 (0.237)

of items, $F(2, 2874) = 5.218$, $p = 0.005$, $\eta^2 = 0.004$. Although models, the number of items, and the interaction effect were statistically significant, only models can explain the significantly large variance of the parameter estimate bias, because the η^2 was the largest and indicated a meaningful difference. It was concluded that discrimination parameter estimates obtained from the model with orthogonality violations between two subdomains had significantly smaller bias than those from the model with orthogonality violations among all subdomains.

For the intercept parameters, there was not a significant effect for the models , $F(1, 1434) = 0.346$, $p = 0.556$, $\eta^2 = 0.001$, an insignificant effect of the number of items, $F(2, 1434) = 1.989$, $p = 0.137$, $\eta^2 = 0.003$, and an insignificant interaction effect between models and the number of items, $F(2, 1434) = 0.528$, $p = 0.590$, $\eta^2 = 0.001$. Therefore, the intercept parameters for the model with violations between two subdomains had no significant differences from the model with violations among all subdomains, and all intercept parameters can be accurately recovered. The number of items also had no significant differences in intercept parameter estimates.

For the theta parameters, there was a significant effect for models, $F(1, 599994) = 4702.485$, $p = 0.001$, $\eta^2 = 0.008$, a significant effect for the number of items, $F(2, 599994) = 11381.942$, $p = 0.001$, $\eta^2 = 0.037$, and a significant interaction effect between models and the number of items, $F(2, 599994) = 84.774$, $p = 0.001$, $\eta^2 = 0.001$. The models, the number of items, and the interaction effects were all statistically significant, but the η^2 was small, and therefore, did not make a significant difference in the parameter estimates. Figure 4.16 is the graphical display of the model and the number of items effects on the item and person parameters.

Table 4.16: Models and Numbers of Items Effects on Item and Person Parameters

	SS	df	MS	F	p	η^2
Discrimination Parameters: a						
Models	6.246	1.000	6.246	470.468	0.001	0.141
Numbers of Items	0.615	2.000	0.308	23.179	0.001	0.016
Models × No. of Items	0.139	2.000	0.069	5.218	0.005	0.004
Error	38.156	2874.000	0.013			
Intercept parameters: d						
Models	0.001	1.000	0.001	0.346	0.556	0.001
Numbers of Items	0.007	2.000	0.004	1.989	0.137	0.003
Model × No. of Items	0.002	2.000	0.001	0.528	0.590	0.001
Error	2.678	1434.000	0.002			
Theta parameters: θ						
Models	245.464	1.000	245.464	4702.458	0.001	0.008
Numbers of Items	1188.252	2.000	594.126	11381.942	0.001	0.037
Model ×No. of Items	8.850	2.000	4.425	84.774	0.001	0.001
Error	31319.093	599994.000	0.052			

Note: The Model variable represents Model 1 and Model 2, and the number of items variable contains 40, 60, and 80.

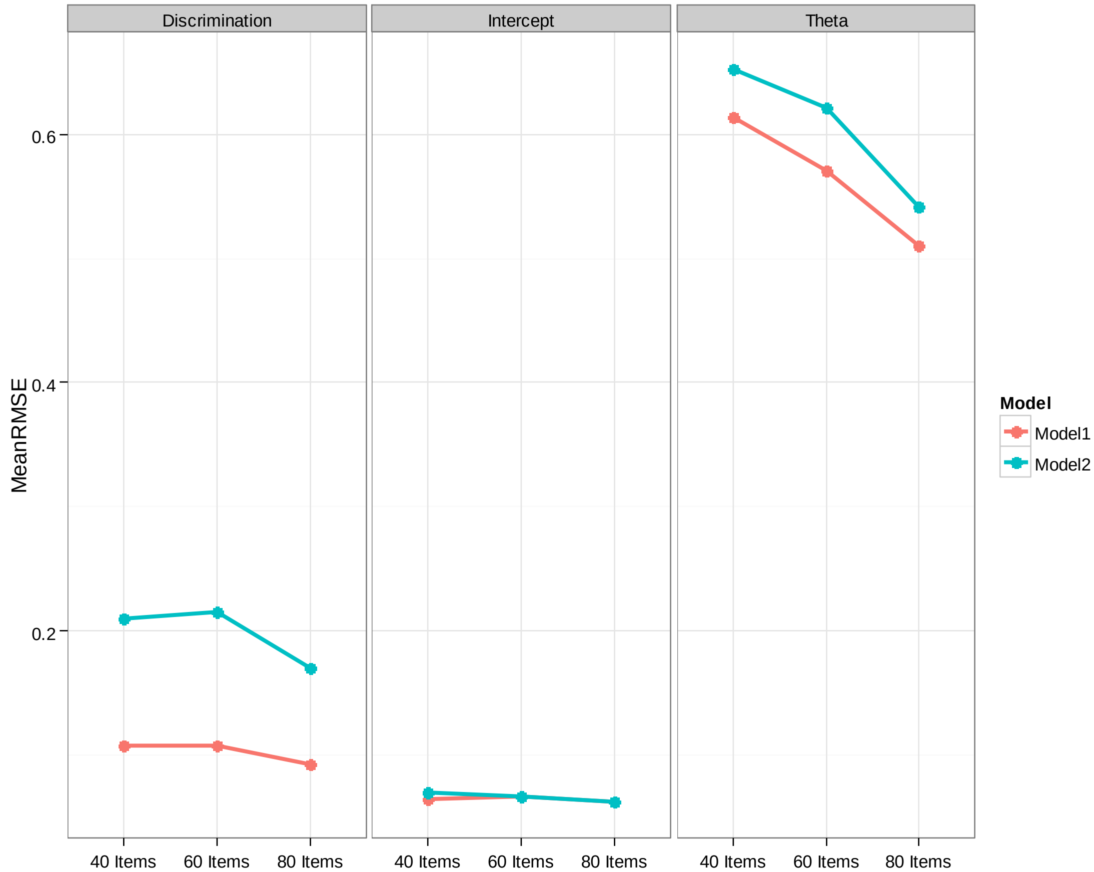

Figure 4.16: Models and Numbers of Items Effects on Item and Person Parameters

4.4 Effects of Specific Models and Dimensions

Model 1 and Model 2 were significantly different in parameter estimate accuracy for the a parameters, but had no significant effects on d and θ parameter estimates. In this section, additional studies were conducted to examine weather the specific models and dimensions were significant predictors for the item and person parameters. The mean bias and the mean RMSE were first reported across each dimension of all specific models, then statistical tests were conducted to examine whether the specific models and dimensions had significant effects on parameter estimates.

4.4.1 Mean Bias and Mean RMSE of a Parameters

Table 4.17 reports the mean bias and the mean RMSE of specific models and dimensions for a parameters. The specific models of Model 1 were slightly different in the mean bias ($\overline{X}_{Bias\&Model11}$= -0.028, $\overline{X}_{Bias\&Model12}$= -0.025, $\overline{X}_{Bias\&Model13}$= -0.045, $\overline{X}_{Bias\&Model14}$= -0.067). Model 13 and Model 14 were slightly higher in the mean bias than Model 11 and Model 12. The mean bias of the primary, the first, and the second dimensions were very close to each other ($\overline{X}_{Bias\&P}$= 0.037, $\overline{X}_{Bias\&S1}$= -0.006, $\overline{X}_{Bias\&S2}$= -0.004). But the mean bias of the third and the fourth dimensions was much higher ($\overline{X}_{Bias\&S3}$= -0.106, $\overline{X}_{Bias\&S4}$= -0.127).

The specific models of Model 1 were very different in terms of mean RMSE ($\overline{X}_{RMSE\&Model11}$= 0.078, $\overline{X}_{RMSE\&Model12}$= 0.084, $\overline{X}_{RMSE\&Model13}$= 0.106, $\overline{X}_{RMSE\&Model14}$= 0.153). The mean RMSE became higher as the orthogonality violation became larger. The mean bias scores of the primary, the first, and the second dimensions were approximately the same ($\overline{X}_{RMSE\&P}$= 0.094, $\overline{X}_{RMSE\&S1}$= 0.074, $\overline{X}_{RMSE\&S2}$= 0.078). But the mean bias scores of the third and the fourth dimensions were much higher ($\overline{X}_{RMSE\&S3}$= 0.130, $\overline{X}_{RMSE\&S4}$= 0.150).

For Model 2, the average mean bias scores for Model 23 and Model 24 were much larger than those for Model 21 and Model 22 ($\overline{X}_{Bias\&Model21}$= -0.036, $\overline{X}_{Bias\&Model22}$= -0.071, $\overline{X}_{Bias\&Model23}$= -0.153, $\overline{X}_{Bias\&Model24}$= -0.260). The mean bias scores of dimensions were very different from one another ($\overline{X}_{Bias\&P}$= 0.143, $\overline{X}_{Bias\&S1}$= -0.154, $\overline{X}_{Bias\&S2}$= -0.224, $\overline{X}_{Bias\&S3}$= -0.182, $\overline{X}_{Bias\&S4}$= -0.231).

The specific models of Model 2 were much different in the mean RMSE ($\overline{X}_{RMSE\&Model11}$= 0.084, $\overline{X}_{RMSE\&Model12}$= 0.129, $\overline{X}_{RMSE\&Model13}$= 0.236, $\overline{X}_{RMSE\&Model14}$= 0.382). The mean RMSEs increased with the growth of the orthogonality violations. The mean RMSE of the dimensions were also very different from each other ($\overline{X}_{RMSE\&P}$= 0.174, $\overline{X}_{RMSE\&S1}$= 0.178, $\overline{X}_{RMSE\&S2}$= 0.240, $\overline{X}_{RMSE\&S3}$= 0.201, $\overline{X}_{RMSE\&S4}$= 0.248). Figure 4.17 is the graphical display of the mean bias and the mean RMSE.

Table 4.17: Mean Bias and Mean RMSE of Specific Models and Dimensions for *a* Parameters

			Dimensions				
			P	S1	S2	S3	S4
$\overline{X}_{Bias}$ (σ)	Model 1	Model 11	0.022 (0.029)	-0.025 (0.027)	-0.030 (0.029)	-0.040 (0.022)	-0.066 (0.051)
		Model 12	0.036 (0.037)	-0.014 (0.026)	-0.006 (0.037)	-0.060 (0.027)	-0.081 (0.056)
		Model 13	0.041 (0.075)	-0.010 (0.030)	-0.005 (0.040)	-0.116 (0.047)	-0.135 (0.059)
		Model 14	0.049 (0.140)	0.025 (0.043)	0.024 (0.056)	-0.208 (0.091)	-0.224 (0.104)
	Model 2	Model 21	0.026 (0.033)	-0.019 (0.024)	-0.061 (0.036)	-0.042 (0.02)	-0.082 (0.059)
		Model 22	0.090 (0.064)	-0.072 (0.031)	-0.147 (0.056)	-0.096 (0.054)	-0.130 (0.064)
		Model 23	0.176 (0.132)	-0.181 (0.073)	-0.267 (0.105)	-0.213 (0.091)	-0.278 (0.110)
		Model 24	0.278 (0.212)	-0.345 (0.130)	-0.422 (0.163)	-0.375 (0.141)	-0.435 (0.166)
$\overline{X}_{RMSE}$ (σ)	Model 1	Model 11	0.067 (0.019)	0.073 (0.024)	0.077 (0.029)	0.076 (0.024)	0.099 (0.056)
		Model 12	0.076 (0.025)	0.071 (0.023)	0.073 (0.028)	0.088 (0.028)	0.110 (0.062)
		Model 13	0.095 (0.046)	0.071 (0.022)	0.076 (0.024)	0.135 (0.046)	0.154 (0.066)
		Model 14	0.138 (0.082)	0.081 (0.021)	0.087 (0.031)	0.220 (0.089)	0.238 (0.106)
	Model 2	Model 21	0.07 (0.021)	0.07 (0.024)	0.092 (0.037)	0.077 (0.022)	0.109 (0.063)
		Model 22	0.114 (0.052)	0.098 (0.033)	0.164 (0.057)	0.119 (0.048)	0.150 (0.069)
		Model 23	0.199 (0.111)	0.194 (0.073)	0.276 (0.104)	0.224 (0.089)	0.289 (0.113)
		Model 24	0.303 (0.185)	0.351 (0.129)	0.429 (0.161)	0.382 (0.139)	0.443 (0.168)

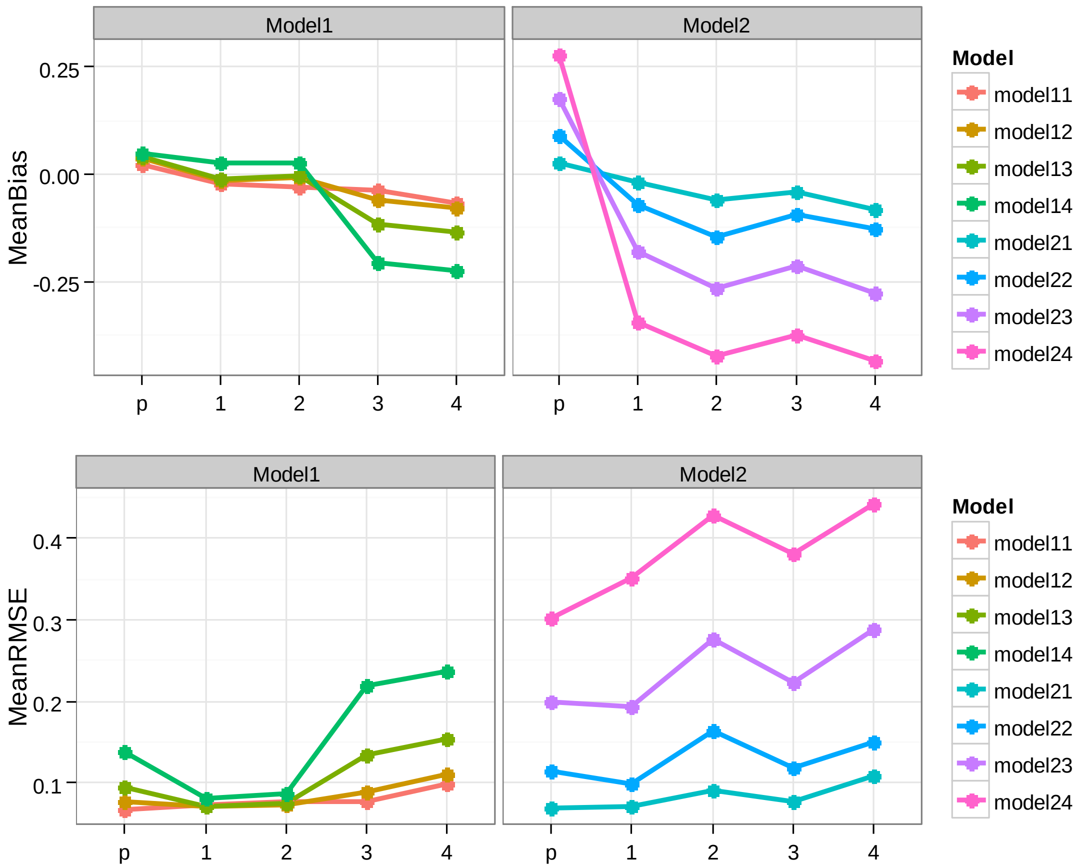

Figure 4.17: Mean Bias and Mean RMSE of Specific Models and Dimensions for *a* Parameters

4.4.2 Mean Bias and Mean RMSE of *d* Parameters

Table 4.18 indicates the mean bias and the mean RMSE of intercept parameters. The mean bias and the mean RMSE for Model 1 were all around zero, indicating the specific models had no effects on intercept estimate accuracy (e.g., $\overline{X}_{Bias\&Model11}$= -0.003, $\overline{X}_{Bias\&Model12}$= -0.007, $\overline{X}_{Bias\&Model13}$= -0.010, $\overline{X}_{Bias\&Model14}$= -0.006; $\overline{X}_{RMSE\&Model11}$= 0.063, $\overline{X}_{RMSE\&Model12}$= 0.064, $\overline{X}_{RMSE\&Model13}$= 0.066, $\overline{X}_{RMSE\&Model14}$= 0.064). The mean bias and the mean RMSE for Model 2 also indicated the same results. Specific models of Model 2 could not affect the intercept estimate accuracy. Figure 4.18 is the graphical display of the mean bias and the mean RMSE of the intercept parameters. As shown by the figure, the mean bias and the mean RMSE were close to each other; differences among specific models were negligible.

Table 4.18: Mean Bias and Mean RMSE of Specific Models for *d* Parameters

		$\overline{X}_{Bias}$ (σ)	$\overline{X}_{RMSE}$ (σ)
Model 1	Model 11	-0.003 (0.042)	0.063 (0.043)
	Model 12	-0.007 (0.045)	0.064 (0.045)
	Model 13	-0.010 (0.045)	0.066 (0.044)
	Model 14	-0.006 (0.041)	0.064 (0.042)
Model 2	Model 21	-0.001 (0.050)	0.065 (0.046)
	Model 22	-0.024 (0.039)	0.066 (0.044)
	Model 23	-0.019 (0.046)	0.069 (0.043)
	Model 24	0.005 (0.040)	0.064 (0.039)

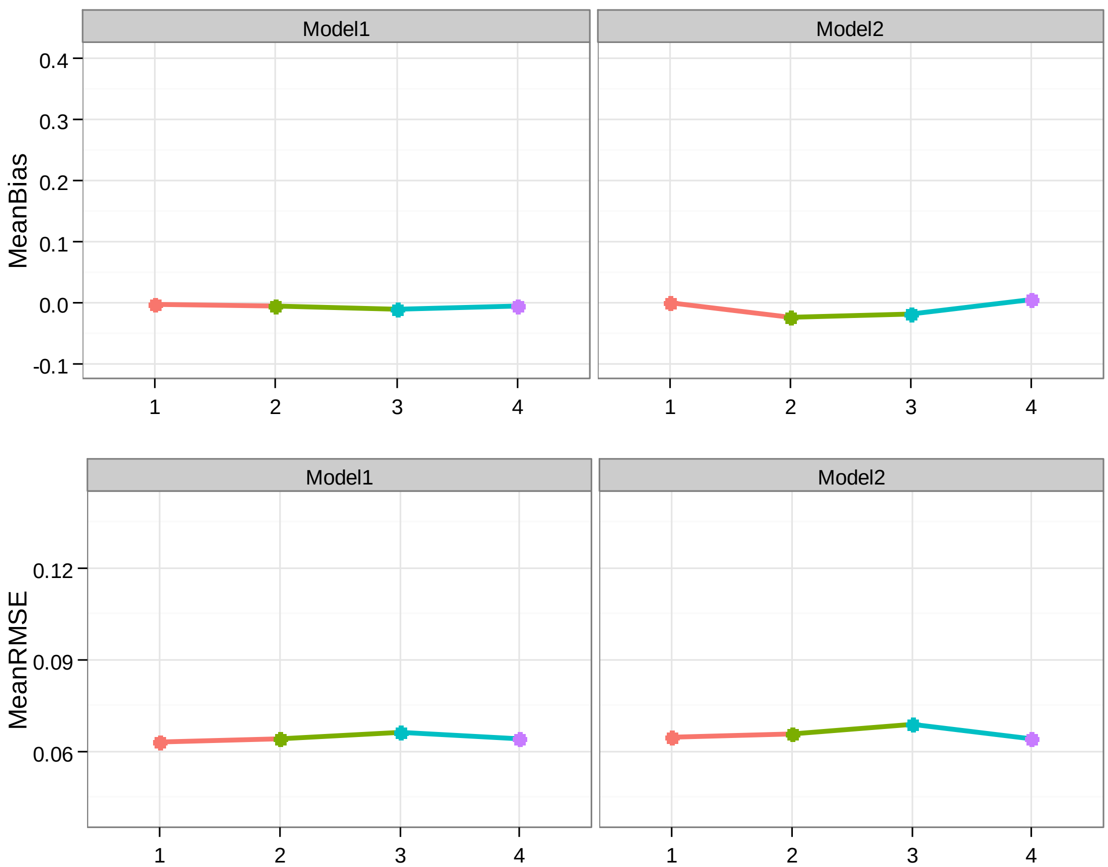

Figure 4.18: Mean Bias and Mean RMSE of Specific Models for *d* Parameters

4.4.3 Mean Bias and Mean RMSE of θ Parameters

Table 4.19 reports the mean bias and the mean RMSE of the specific models and dimensions for the θparameters. For Model 1, t he mean bias across specific models and dimensions had small differences. They were all centered around zero (e.g., $\overline{X}_{BiasP\&Model11}$= -0.003, $\overline{X}_{BiasS1\&Model11}$= 0.006, $\overline{X}_{BiasS2\&Model11}$= -0.011, $\overline{X}_{BiasS3\&Model11}$= 0.008, $\overline{X}_{BiasS4\&Model11}$= 0.002). The mean RMSE of specific models of Model 2 had few differences ($\overline{X}_{RMSE\&Model11}$= 0.556, $\overline{X}_{RMSE\&Model11}$= 0.560, $\overline{X}_{RMSE\&Model11}$= 0.564, $\overline{X}_{RMSE\&Model11}$= 0.580). However, the dimensions of the specific models indicated some differences ($\overline{X}_{RMSE\&P}$= 0.443, $\overline{X}_{RMSE\&S1}$= 0.553, $\overline{X}_{RMSE\&S2}$= 0.589, $\overline{X}_{RMSE\&S3}$= 0.601, $\overline{X}_{RMSE\&S4}$= 0.639).

For Model 2, the mean bias scores across the specific models and the dimensions were almost the same. They were also centered around zero. The mean RMSE of the specific models differed a little bit ($\overline{X}_{RMSE\&Model11}$= 0.561, $\overline{X}_{RMSE\&Model11}$= 0.577, $\overline{X}_{RMSE\&Model11}$= 0.617, $\overline{X}_{RMSE\&Model11}$= 0.666). The mean RMSE of the dimensions indicated bigger differences ($\overline{X}_{RMSE\&P}$= 0.481, $\overline{X}_{RMSE\&S1}$= 0.596, $\overline{X}_{RMSE\&S2}$= 0.638, $\overline{X}_{RMSE\&S3}$= 0.637, $\overline{X}_{RMSE\&S4}$= 0.675). Figure 4.19 is the graphical display of the mean bias and the mean RMSE across all specific models and dimensions.

Table 4.19: Mean Bias and Mean RMSE of Specific Models and Dimensions for θ Parameters

			Dimensions				
			P	S1	S2	S3	S4
$\overline{X}_{Bias}$ (σ)	Model 1	Model 11	-0.003 (0.313)	0.006 (0.374)	-0.011 (0.419)	0.008 (0.448)	0.002 (0.456)
		Model 12	0.001 (0.321)	0.001 (0.384)	-0.001 (0.428)	0.009 (0.452)	-0.002 (0.473)
		Model 13	0.002 (0.337)	0.023 (0.376)	0.001 (0.425)	-0.006 (0.464)	0.006 (0.482)
		Model 14	0.002 (0.381)	0.001 (0.391)	0.001 (0.436)	-0.001 (0.509)	-0.013 (0.523)
	Model 2	Model 21	0.001 (0.327)	0.017 (0.386)	-0.018 (0.422)	-0.014 (0.457)	0.003 (0.467)
		Model 22	0.012 (0.372)	0.008 (0.406)	-0.001 (0.448)	0.001 (0.482)	0.009 (0.497)
		Model 23	0.012 (0.460)	0.020 (0.453)	0.005 (0.504)	0.001 (0.545)	0.008 (0.562)
		Model 24	0.003 (0.546)	-0.004 (0.521)	-0.006 (0.590)	-0.006 (0.626)	-0.006 (0.653)
$\overline{X}_{RMSE}$ (σ)	Model 1	Model 11	0.433 (0.142)	0.552 (0.171)	0.587 (0.193)	0.587 (0.225)	0.621 (0.205)
		Model 12	0.436 (0.146)	0.554 (0.181)	0.589 (0.198)	0.590 (0.227)	0.631 (0.218)
		Model 13	0.442 (0.155)	0.551 (0.172)	0.588 (0.193)	0.600 (0.234)	0.638 (0.222)
		Model 14	0.462 (0.182)	0.555 (0.180)	0.592 (0.196)	0.626 (0.262)	0.664 (0.249)
	Model 2	Model 21	0.439 (0.151)	0.556 (0.181)	0.592 (0.192)	0.591 (0.229)	0.629 (0.210)
		Model 22	0.453 (0.177)	0.570 (0.192)	0.611 (0.209)	0.606 (0.243)	0.646 (0.230)
		Model 23	0.493 (0.232)	0.605 (0.215)	0.648 (0.238)	0.649 (0.277)	0.688 (0.267)
		Model 24	0.537 (0.285)	0.654 (0.250)	0.701 (0.287)	0.701 (0.323)	0.738 (0.323)

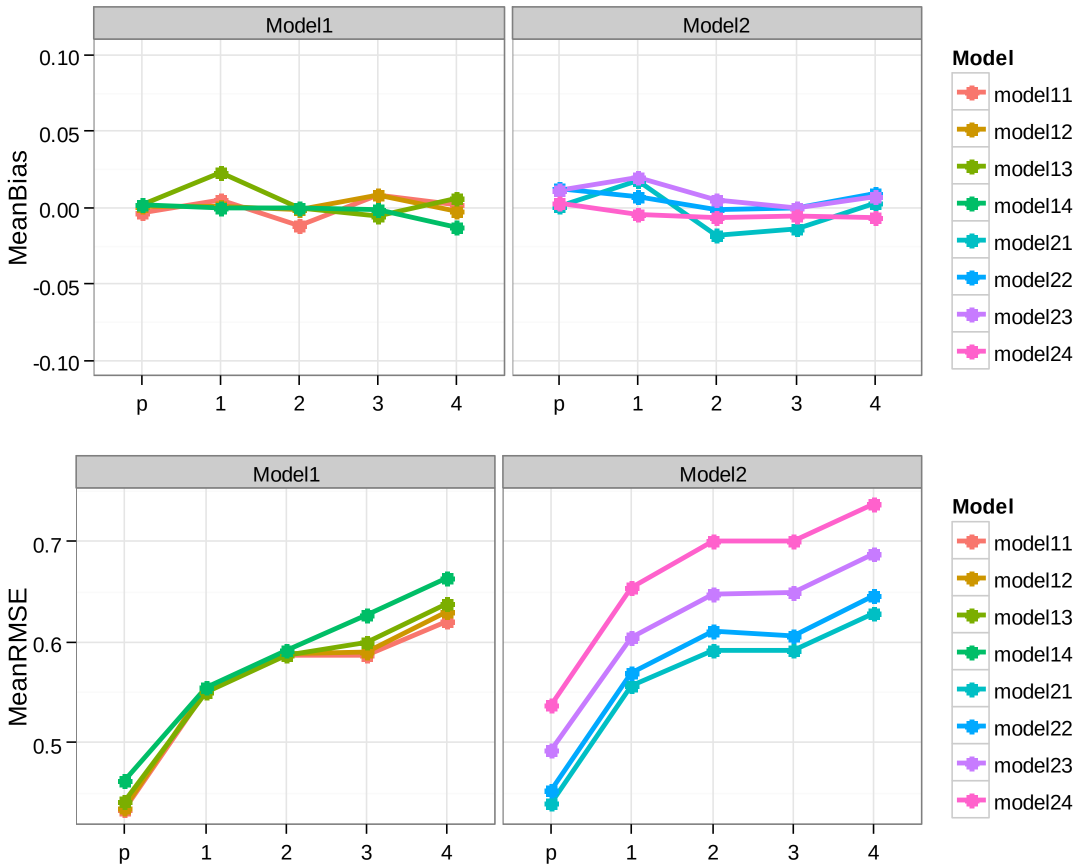

Figure 4.19: Mean Bias and Mean RMSE of Specific Models and Dimensions for θ Parameters

4.4.4 Specific Model and Dimension Effects on a, d, and θ Parameters

Table 4.20 presents the statistical test results of specific models and dimensions on a parameters. For Model 1, there was significant effect for dimensions, $F(4, 1420) = 86.334$, $p = 0.001$, $\eta^2 = 0.196$, a significant effect for specific models, $F(3, 1420) = 125.014$, $p = 0.001$, $\eta^2 = 0.209$, and a significant interaction effect between the dimensions and the specific models for the parameter estimate bias, $F(12, 1420) = 1.989$, $p = 0.001$, $\eta^2 = 0.121$. There was a significant main effect for dimensions and specific models, and the mean RMSE of dimensions also depended on the specific models.

For Model 2, there was a significant effect on dimensions, $F(4, 1420) = 30.432$, $p = 0.001$, $\eta^2 = 0.079$, a significant effect on specific models, $F(3, 1420) = 431.660$, $p = 0.001$, $\eta^2 = 0.477$, and a significant interaction effect between specific models and dimensions, $F(12, 1420) = 3.693$, $p = 0.001$, $\eta^2 = 0.030$. Even though all of them are statistically significant, only specific models explained a significant difference. Table 4.21 illustrate the results of the additional statistical test examining which specific models are significantly different from one another. The results indicated that all of the specific models were significant from one another.

Table 4.20: Specific Models and Dimensions Effects for *a* Parameters

	SS	df	MS	F	p	η^2
Model 1:						
Dimensions	0.842	4.000	0.211	86.334	0.001	0.196
Specific Models	0.915	3.000	0.305	125.014	0.001	0.209
Dimensions × Specific Models	0.477	12.000	0.040	16.279	0.001	0.121
Error	3.464	1420.000	0.002			
Model 2:						
Dimensions	1.310	4.000	0.328	30.432	0.001	0.079
Specific Models	13.939	3.000	4.646	431.660	0.001	0.477
Dimensions × Specific Models	0.477	12.000	0.040	3.693	0.001	0.030
Error	15.284	1420.000	0.011			

Table 4.21: Post-Hoc Test for the Specific Models of Model 2

	difference	95% CI	p
M22-M21	0.045	0.025 - 0.065	<0.001
M23-M21	0.144	0.124 - 0.164	<0.001
M24-M21	0.274	0.254 - 0.293	<0.001
M23-M22	0.099	0.079 - 0.119	<0.001
M24-M22	0.229	0.209 - 0.248	<0.001
M24-M23	0.130	0.110 - 0.149	<0.001

Table 4.22 reports the specific model effects on the intercept parameters. As indicated by the table, the specific models of Model 1 had no significant effect on the intercept parameters, $F(3, 716) = 0.137$, $p = 0.938$, $\eta^2 = 0.001$. Similarly, the specific models of Model 2 also had no significant effect on the intercept parameters, $F(3, 716) = 0.458$, $p = 0.711$, $\eta^2 = 0.002$.

Table 4.22: Specific Model Effects for d Parameters

	SS	df	MS	F	p	η^2
Model 1:						
Specific Models	0.001	3	0.001	0.137	0.938	0.001
Error	1.348	716	0.002			
Model 2:						
Specific Models	0.003	3	0.001	0.458	0.711	0.002
Error	1.336	716	0.002			

Table 4.23 demonstrates a statistical test of dimensions and specific models on θparameters. For Model 1, there was a significant effect of dimensions, $F(4, 299980) = 8312.371$, $p = 0.001$, $\eta^2 = 0.100$, a significant effect of the specific models, $F(3, 299980) = 205.176$, $p = 0.001$, $\eta^2 = 0.002$, and a significant interaction effect, $F(12, 299980) = 26.955$, $p = 0.001$, $\eta^2 = 0.001$. Only dimension can account for a significant variance of the mean RMSE because the η^2 is 0.100.

For Model 2, there was a significant effect on dimensions, $F(4, 299980) = 5886.697$, $p = 0.001$, $\eta^2 = 0.073$, a significant effect on the specific models, $F(3, 299980) = 2837.710$, $p = 0.001$, $\eta^2 = 0.028$, and a significant interaction effect, $F(12, 299980) = 2.253$, $p = 0.008$, $\eta^2 = 0.001$. Since all of the η^2 are small, the dimensions and the specific models were actually not significant predictors of the mean RMSE.

Table 4.24 reports further analysis of the dimensions for Model 1. Results indicated that all of the dimensions were significant from each other (p <0.001). The mean differences between the primay dimension and subdomains were larger ($\overline{X}_{S1} - \overline{X}_P = 0.110$, $\overline{X}_{S2} - \overline{X}_P = 0.146$, $\overline{X}_{S3} - \overline{X}_P = 0.157$, $\overline{X}_{S4} - \overline{X}_P = 0.195$). All of the other mean differences were small, even though the differences are significant.

Table 4.23: Specific Models and Dimensions Effects on θ Parameters

	SS	df	MS	F	p	η^2
Model 1:						
Dimensions	1331.986	4	332.997	8312.371	.001	.100
Specific Models	24.658	3	8.219	205.176	.001	.002
Dimensions × Specifc Models	12.958	12	1.080	26.955	.001	.001
Error	12017.309	299980	.040			
Model 2:						
Dimensions	1356.453	4	339.113	5886.697	.001	.073
Specific Models	490.413	3	163.471	2837.710	.001	.028
Dimensions × Specific Models	1.558	12	.130	2.253	.008	.001
Error	17280.860	299980	.058			

Table 4.24: Post-Hoc Test on Dimensions of Model 1

	Differences	95%CI	p
S1 - P	0.110	0.106 - 0.113	<0.001
S2 - P	0.146	0.142 - 0.149	<0.001
S3 - P	0.157	0.154 - 0.161	<0.001
S4 - P	0.195	0.192 - 0.198	<0.001
S1 - S2	0.036	0.033 - 0.039	<0.001
S1 - S3	0.048	0.045 - 0.051	<0.001
S1 - S4	0.086	0.082 - 0.089	<0.001
S2 - S3	0.012	0.009 - 0.015	<0.001
S2 - S4	0.050	0.046 - 0.053	<0.001
S3 - S4	0.038	0.035 - 0.041	<0.001

4.5 Empirical Results from the Real Data

The purpose of this section is to describe the possible bias of real data when implementing a bifactor analysis. As stated in the method section, a bifactor analysis was conducted to obtain the item and person parameter estimates. The target oblique rotation method was subsequently implemented to explore the possible correlations among the subdomains. Based on the simulation study, the possible bias scores were then discussed for the item and person parameter estimates.

Table 4.25 describes the possible correlations among all the subdomains. As shown in the table, correlations existed among all the subdomains, $r_{Data,Number} = 0.235$, $r_{Geometry,Number} = 0.165$, $r_{Algebra,Number} = 0.251$, $r_{Geometry,Data} = 0.133$,$r_{Algebra,Data} = 0.253$, $r_{Albebra,Geometry} =$

0.309. Model 1 in the simulation study examined correlations only between two subdomains, while Model 2 studied correlations among all subdomains. Because correlations occurred among all the subdomains in the real data, it fell into Model 2 category. In addition, because all of the correlations in the real data are roughly between 0.100 and 0.300, the bias is therefore interpreted based on Model 22 with 80 items.

Table 4.25: Possible Correlations Among the Subdomains of the Real Data

	Math	Number	Data	Geometry	Algebra
Math	1.000				
Number	0.000	1.000			
Data	0.000	0.235	1.000		
Geometry	0.000	0.165	0.133	1.000	
Algebra	0.000	0.251	0.253	0.309	1.000

4.5.1 Possible Bias and RMSE of *a* and *d* Parameters

Based on Model 22 in Table 4.9, the intercept parameter estimates from the real data should have very few differences in comparison to the true parameter. All intercept parameters could be accurately recovered because the simulation data suggested both a small mean bias and the variability between the estimated and the true parameters. There may be a few outliers, but the overall bias should be small. The mean bias for d parameters should be around -0.021, while the mean RMSE should be around 0.063, as shown by Table 4.14. The parameter estimates, $\overline{X}$ =1.413, σ =0.815, and quartiles statistics, are trustworthy.

For the discrimination parameter, the bias between the estimated and the true parameter should be small as indicated by Model 22 in Table 4.9. The parameter for the primary dimension, Math, should be slightly overestimated, and all the other subdomain parameters, Number, Data, Geometry, and Algebra, should be underestimated (see Model 22). The mean bias for a parameter should be around -0.014, and the mean RMSE should be around 0.104, as indicated by Table 4.13. There should not be a significant difference between the estimates and the true parameters, but higher correlations can lead to higher RMSE.

Table 4.26: Item Parameter Estimates of the Real Data

	$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Intercept	1.413	0.815	-0.326	0.870	1.361	1.980	3.292
Math	1.204	0.333	0.316	1.003	1.161	1.388	2.412
Number	0.190	0.605	-1.396	-0.140	0.198	0.626	1.397
Data	-0.085	1.310	-2.955	-0.389	-0.058	0.629	2.650
Geometry	0.204	0.557	-0.712	-0.168	0.286	0.485	1.308
Algebra	-0.156	0.909	-2.359	-0.551	-0.019	0.255	2.169

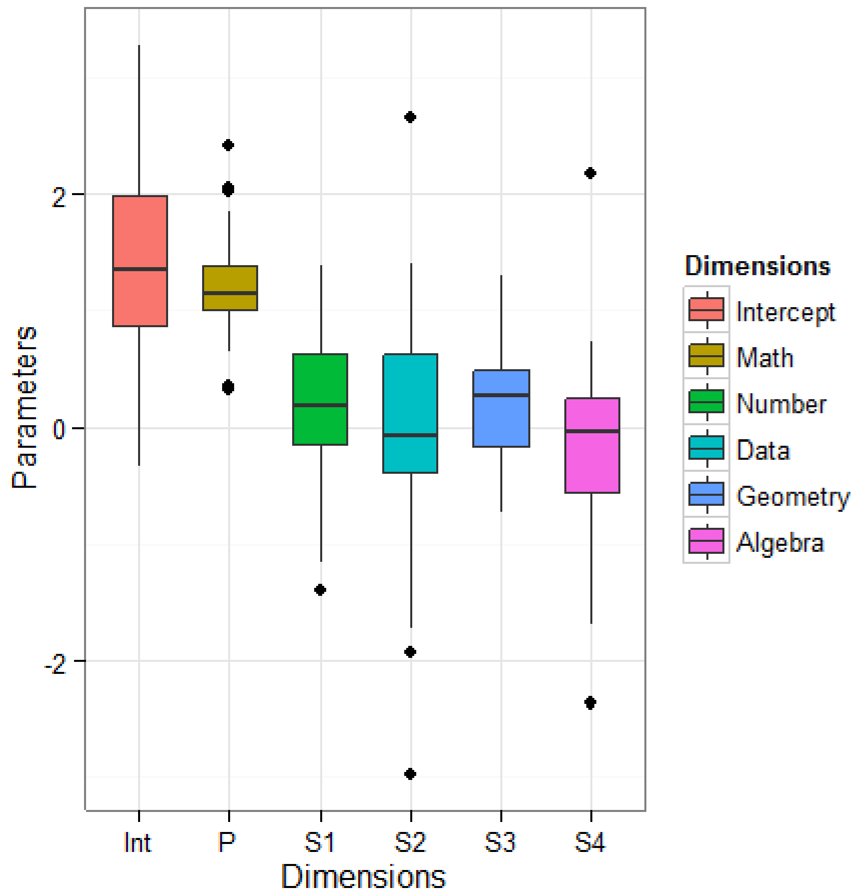

Figure 4.20: Item Parameter Estimates of the Real Data

4.5.2 Possible Bias and RMSE of θ Parameters

Based on Model 22 in Table 4.12, the mean and median of the estimates of the real data should be approximately the same as the true parameters; therefore, the mean of math, number, data, geometry, and algebra illustrated by Table 4.27 should be close to the mean of true parameters. The standard deviation for the primary dimension, Math, should be the one most closely aligned to the true standard deviation compared with other subdomains; thus, range of parameters for the first dimension can be better recovered. The range of the subdomain dimension should be even narrower than the first dimensions. The mean bias for the θ parameters should be around 0.005, and the mean RMSE should be around 0.517, as indicated by 4.15.

Table 4.27: Person Parameter Estimates of the Real Data

	$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
Math	-0.003	0.954	-2.791	-0.709	-0.047	0.689	2.532
Number	0.001	0.754	-2.205	-0.558	0.037	0.526	2.114
Data	-0.002	0.830	-2.096	-0.616	-0.042	0.562	2.481
Geometry	0.001	0.600	-2.487	-0.310	0.077	0.338	1.818
Algebra	-0.001	0.828	-2.035	-0.767	0.096	0.683	2.266

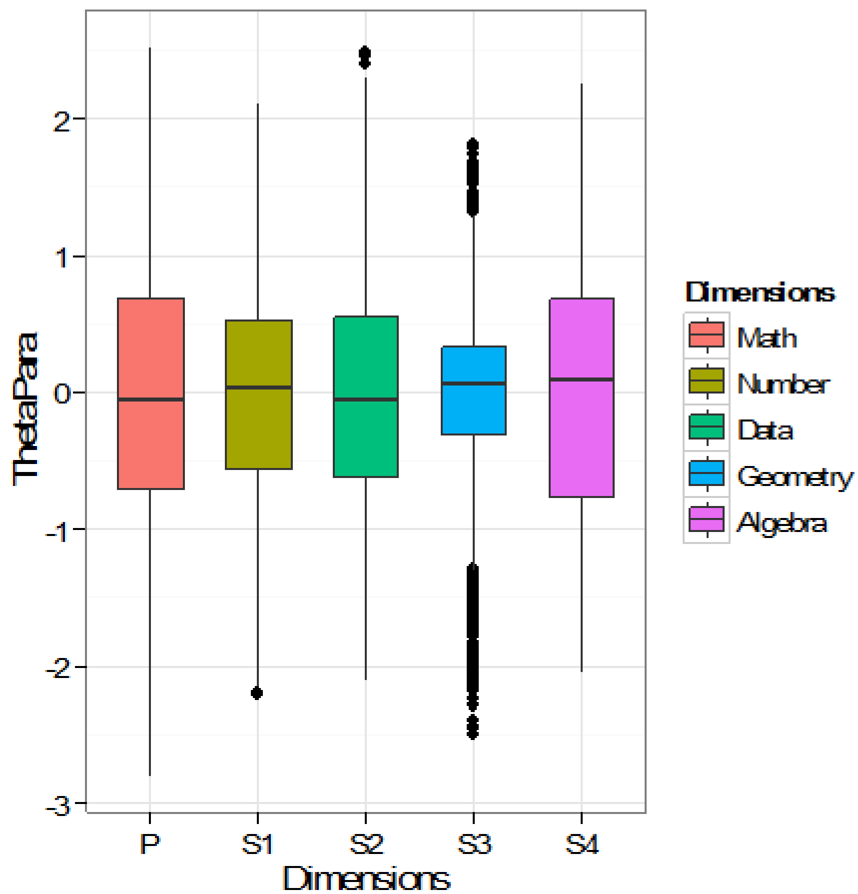

Figure 4.21: Person Parameter Estimates of the Real Data

Chapter 5

Conclusion and Discussion

This chapter summarizes and discusses findings of the present study. First, a brief summary of the study is made, then findings, conclusions, implications, limitations and directions for future research are summarized and discussed. The findings are presented in the order of parameter estimate bias when the orthogonality violations are between two subdomains (i.e., Model 1), parameter estimate bias when the orthogonality violations are among all subdomains (i.e., Model 2), effects of models (i.e., Model 1 and Model 2) and numbers of items, and effects of specific models with different levels of orthogonality violation and dimensions.

5.1 Summary of the Study

The structure of educational and psychological assessments normally consists of a measured construct and several related subdomains for specifying the measured construct. Data collected from those assessments are multidimensional because the measured construct is defined and measured by a set of related subdomains. To accurately describe or make inferences about such data, a model needs to capture the structure with both the primary and the related specific factors.

A bifactor model can capture the structure of educational and psychological assessments with one primary dimension and a set of related subdomains. It assumes that all items measuring the primary construct explains the variance among participants, and that the specific two or more subdomains explain the variance not accounted for by the primary construct.The bifactor model has been increasingly applied to empirical data from achievement tests and multiple-domain psychological instruments, mainly because it can control the possible distortion of multidimensional data and provide information for accountability and diagnosis purposes.

However, a limitation of the bifactor model is the orthogonal constraint that there are no correlations among the subdomains. This assumption requires that items be written perfectly to measure the primary factor and only one subdomain. In reality, it is difficult to meet the assumption because items are more or less flawed in practice. It is rarely possible to write perfect items to measure one primary construct and only one subdomain. To force correlated factors to be orthogonal can result in a loss of information and can lead to distorted and untrustworthy item parameter estimates in the bifactor solutions. Little has been investigated regarding the parameter estimate bias if the orthogonality assumption is violated. Precision of parameter estimates is an important issue in any assessment since parameter estimates are considered to be decisive criteria for finalizing item performance and respondents' ability levels. The present research, therefore, investigated the parameter estimate bias of different levels of orthogonality of the bifactor model.

Since the orthogonality violation can't be controlled and the true parameters are unknown in real data, an extensive series of simulated data were generated to evaluate the parameter estimate bias. The simulated data were dichotomously-scored. The manipulated factors included violations between two subdomains or among all subdomains; four levels of orthogonality violation classified as trivial, small, medium, and large; and a total number of items. The fixed factors included a sample size of 5,000 participants, five dimensions with one primary and four related subdomains, and an equal number of items in each construct. A total of 24 cases were examined in the present study ($2 \times 4 \times 3$): violations between two subdomains or among all subdomains (i.e., 2 conditions); four levels of orthogonality violation (i.e., 4 conditions); and the total number of 40, 60, and 80 items (i.e., 3 conditions). In order to minimize the sample variance and increase the power to detect the effects of interest, 200 replications were generated within each case comparing the differences between the estimated and the true parameter estimates. An empirical study with a real data example was examined later to explore the possible bias.

Validity of a bifactor program for data generation written by *R* 2.15 was tested first. Data without violations were simulated to test the true parameters' recovery. The true item and person parameters were saved in an Excel file, and then the bifactor analysis was conducted using the simulated data to test the recovery of the parameter estimates. Results indicated that the intercept and the discrimination parameters were accurately recovered because the bias scores between the estimated and the true parameters were centered around zero and the variability of the bias scores was very small. These results showed that the bifactor program for data generation was valid and thus trustworthy.

Data from the 24 cases were subsequently generated and analyzed. Common evaluation criteria such as bias, RMSE, and SE were computed, but bias and RMSE were further studied to examine the parameter estimate recovery. Descriptive analysis such as mean, standard deviation, and quartiles was first conducted to describe the distribution of the bias scores between the estimated and the true parameters. Mean bias and mean RMSE were then computed to describe the differences across different models, numbers of items, and dimensions. ANOVA was finally performed to test if there were significant differences in the mean RMSE across the different models, numbers of items, and dimensions.

5.1.1 Summary of Parameter Estimate Bias of Model 1

Model 1 described the orthogonality violations between the third and the fourth subdomains of the proposed bifactor model. Four levels of violations such as trial, small, medium, and large were denoted as specific models and labeled as Model 11, Model 12, Model 13, and Model 14 respectively. Bias scores between the estimated and the true parameters were computed to examine the paramter estimate difference. Results indicated the orthogonality violations between two subdomains did not affect the intercept parameter estimates. The intercept estimates were nearly identical because the mean bias scores were centered at zero, and the range of variability was small. There were a few outliers of the bias scores for each specific model, but most of the intercept estimates were almost the same as the true parameters.

The recovery of discrimination parameters of the first and the second subdomains was not affected by the orthogonality violations either. The parameter estimates of the first and second subdomains were approximately equivalent to the true parameters because the bias scores and the range of variance were small. However, violations between the third and fourth subdomains did affect the parameter estimates of the involved subdomains and the primary dimension. The bias scores between the estimated and true parameters increased as the levels of orthogonality violations were more severe between the third and fourth subdomains. The parameter estimates of Model 11 were closest to the true parameters, and the parameter estimates of Model 14 were the farthest from the true parameters. Additionally, the estimates of the subdomains were underestimated because all bias scores were below zero.

Violations between the subdomains had little influence on the theta estimates. The distribution of the theta estimates across all specific models was very similar. The mean bias had few differences from that of the true parameters, however, the range of variability for each dimension was widespread. Of all the parameter estimates, the estimates of the primary dimension were the closest to the true parameters, and thus the most trustworthy.

5.1.2 Summary of Parameter Estimate Bias of Model 2

Model 2 defined the orthogonality violations among all subdomains of the proposed bifactor model. The four levels of violations such as trial, small, medium, and large were also denoted as specific models but labeled as Model 21, Model 22, Model 23, and Model 24 respectively. Bias scores between the estimated and the true parameters were computed as well. For intercept parameters, the orthogonality violations among all subdomains did not influence the intercept parameter recovery either. There were a few outliers, but overall the accuracy of intercept parameters were not influenced by the magnitude of the orthogonality violations.

For discrimination parameters, the differences between the estimated and true parameters became more distorted as the orthogonality violations became larger. The parameter estimates of Model 21 were the least distorted among all the specific models, and the estimates of Model 24 were the most distorted. Even though the correlation violations existed in the subdomains, the estimates of the primary dimensions were also affected. The estimates of all subdomains were underestimated because the bias scores were below zero.

The orthogonality violations had little effect on theta parameter estimates. The distribution of the theta parameters across all specific models and dimensions had a similar tendency. The mean bias scores of the theta estimates were centered on zero, but the bias scores varied greatly. Of all the theta esimates across different dimensions, the estimates of the primary dimension were the closest to the true parameters.

5.1.3 Summary of Effects of Models and Numbers of Items

Models refers to Model 1 violations between the third and fourth subdomains, and Model 2 violations among all subdomains. Numbers of Items were 40, 60, and 80 items. The mean bias and the mean RMSE were computed to examine the differences across the two models and the different numbers of items. ANOVA was then implemented to examine the magnitude of the effect using the mean RMSE. For the discrimination estimates, both the mean bias and the mean RMSE of Model 2 were larger than those of Model 1. The ANOVA result indicated that the parameter estimate bias of Model 2 was significantly larger than the bias of Model 1, but the number of items had no significant influence on parameter estimate bias. There were no significant interaction effects between models and numbers of items. For intercept parameters, the mean bias and the mean RMSE of Model 1 and Model 2 did not differ greatly from each other. The ANOVA result also indicated no significant difference in the mean RMSE between Model 1 and Model 2. The number of items had no significant influence in the mean intercept parameter estimate bias as well. There were no significant interaction effects between models and numbers of items. For theta parameters, both the models and numbers of items had no significant differences on the mean RMSE. There were no significant interaction effects in the mean RMSE either.

5.1.4 Summary of Effects of Specific Models and Dimensions

Specific models characterize the four levels of the orthogonality violations, and dimensions refer to the primary dimension and the related four subdomains of the proposed bifactor model. The mean bias and the mean RMSE were computed to examine the differences across the specific models and the different dimensions. Similarly, ANOVA was then implemented to examine the magnitude of the effect using the mean RMSE. For Model 1, ANOVA results revealed that both specific models and dimensions had significant main effects on the discrimination estimate bias. Parameter estimate bias of specific models with severe orthogonality violations were significantly larger from those of specific models with less severe violations. Likewise, parameter estimate bias of the third and fourth subdomains were also significant larger than those of the first and second subdomains. There were significant interaction effects between the specific models and dimensions. For the intercept parameters, parameter estimate bias was not significantly different across all specific models. For theta parameters, specific models had no significant difference in the estimate bias, however, dimensions did indicate significant differences in the estimate bias. There were also no significant interaction effects between specific models and dimensions.

For Model 2, ANOVA results indicated that specific models had significantly effect on the discrimination parameter estimate bias. Parameter estimates of specific models with severe violations were more distorted than those of specific models with less severe violations. Dimensions were not a significant predictor for the parameter estimate bias. There were no significant effects between specific models and dimensions. For the intercept and theta parameters of Model 2, both specific models and dimensions were not significant predictors of parameter estimate bias. There were also no significant interaction effects between the specific models and dimensions.

5.2 Conclusion of the Study

The levels of orthogonality violation can only result in significant parameter bias of the discrimination parameters. The higher the orthogonality violations are, the more distorted the parameter estimates will be. Parameters obtained from a bifactor model with orthogonality violations between certain subdomains are more accurate than those with orthogonality violations among all subdomains because the violations between certain subdomains only affect the parameter estimates of the involved subdomains and the primary dimension. The parameter estimates of the other subdomains are not affected. Parameter estimates of the subdomains with the orthogonality violations are all underestimated. The greater the magnitude of the orthogonality violation, the greater extent to which the discrimination parameters are underestimated. Numbers of items of 40, 60, and 80 have no significant influence on parameter estimates.

Intercept parameters can be accurately recovered regardless of the orthogonality violations and numbers of items. The mean bias of intercept parameters are around zero, and the parameter estimates do not vary much from the true parameters. There might be a few estimates which differ greatly from the true parameters, but overall the intercept parameter estimates are trustworthy.

The theta parameter estimates, however, are not trustworthy. Theta parameter estimates fluctuate greatly from the true parameters even though the validity model with no orthogonality violations can not accurately recover the true parameters (see Table A.1 and Figure A.1). Of all the theta parameter estimates, those of the primary dimension are relatively trustworthy because the bias variability is relatively small compared with those of other subdomains.

The large discrepancies between the estimated and the true theta parameters may be the result of estimate error. They may also suggest that the bifactor model cannot estimate larger theta parameters, such as values beyond 3.000. If the true person parameters had indices beyond 3.500, while the estimated parameters could only estimate parameters below 2.500, then the bias would be large.

5.3 Implications for Testing Practice

The findings from this study make several contributions. First, the results add substantially to our understanding of bifactor parameter estimate accuracy. The results indicate that different levels of orthogonality violations among subdomains result in different levels of parameter distortion; therefore, the ability and item performance estimates may not be trustworthy at all. Inaccurate ability estimates can lead to severe consequences by misjudging students' ability levels in both the primary and subdomains and by misjudging item performance. Recognizing the potential parameter inaccuracy can help the actual application of the bifactor model and interpretation of the results. Researchers or practitioners may first test the possible correlations among the subdomains before applying the bifactor model.

Second, the study reveals the appropriate number of items for the bifactor analysis. For the present study, the number of items between 40 and 80 has no significant effect on parameter estimates of a bifactor model.

Finally, this research serves as a basis for future studies on the bifactor analysis with oblique rotation. Oblique rotation should theoretically render a more accurate and useful replicable solution. If the factors are uncorrelated, orthogonal and oblique rotation are likely to produce identical results. To ensure the precision of parameter estimates, exploratory bifactor analysis with oblique rotation can be implemented first, and then the correlations among the factors can be examined. However, exploratory bifactor analysis can only indicate to which parameter estimates might be distorted; it cannot solve the problem fundamentally. Therefore, future research on a bifactor item response model with oblique rotation should theoretically render a more accurate, and perhaps more reproducible, solution.

5.4 Limitations and Directions for Future Research

Generalization of simulation results is limited. The results of the simulation studies were based on a bifactor structure with one primary dimension and four subdomains, the number of items of 40, 60, and 80, and 5,000 participants. It is difficult to determine weather the results are generalizable to a bifactor model with different dimensions, different numbers of items, and different participants. If there are six, seven, or more dimensions of a bifactor model, will the parameter estimates follow the same pattern? The number of items has no significant effect on parameter estimates based on the simulation results, but what if the number of items is 20, 30, or 120? And what if the participants in the bifactor model are 500, 1,000, or 2,000? Practitioners should be cautious when interpreting the results of real data based on the simulated results.

Additionally, this study demonstrated the possible bias of parameter estimates for a bifactor model when violating the orthogonal assumption. For real data analysis, a target rotation matrix method was suggested to explore the possible correlations among the subdomains using Mplus. However, no published studies have been conducted on the correlation recovery of target rotation matrix. Future studies should focus on the correlation recovery of the target rotation matrix method. If the target rotation matrix can actually recover the correlations among the subdomains, the bias of real data can be interpreted based on the simulated results.

Finally, among the theta parameter estimates, the values of the primary dimension were closest to the true parameters. All estimates of the subdomains had larger discrepancies between the estimates and the true parameters. Even for the ideal model where there were no correlation violations, the theta parameter estimates could not be accurately recovered. Because theta estimates are important criteria for deciding students' ability level, accuracy of theta parameters is crucial. Future studies should focus on improving the accuracy of theta estimates.

References

Asparouhov, T., & Muthen, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, *27*(6), 395–414.

Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the dispositional hope scale. *Psychological Assessment*, *20*(3), 310–315.

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to assess the dimensionality of the hong psychological reactance scale. *Educational and Psychological Measurement*, *71*(1), 170–185.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2004). CEFA: comprehensive exploratory factor analysis, version 2.00 [Computer software and manual]. *The Ohio State University, Columbus*.

Cai, L. (2010a). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cai, L. (2010b). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis.

Psychological Methods, *16*(3), 221–248.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225.

De la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, *35*(4), 296–316.

DeMars, C. E. (2006). Application of the Bi-Factor multidimensional item response theory model to Testlet-Based tests. *Journal of educational measurement*, *43*(2), 145–168.

Ebesutani, C., Smith, A., Bernstein, A., Chorpita, B. F., Higa-McMillan, C., & Nakamura, B. (2011). A bifactor model of negative affectivity: Fear and distress components among younger and older youth. *Psychological Assessment*, *23*(3), 679–691.

Fukuhara, H. (2009). *A differential item functioning model for testlet-based items using a bi-factor multidimensional item response theory model: A bayesian approach*. Doctoral dissertation, retrieved from http://search.proquest.com/docview/304883016?accountid=14556.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., … Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler adult intelligence scale (WAIS-III). *Psychological Assessment*, *23*(1), 143–152.

Hallquist, M. (2011). Mplusautomation: Automating mplus model estimation and interpretation. *R package version 0.5 [online]. Available at http://cran. r-project. org/web/packages/MplusAutomation/index. html*.

Højsgaard, S. (2012). The doby package. *The Newsletter of the R Project Volume 6/2, May 2006*, *1*(1), 47.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data an illustration with the state metacognitive inventory. *Educational and Psychological Measurement*, *68*(4), 695–709.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 395–416.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, *76*(4), 537–549.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

Kell. (2012). kobe: Tools for providing advice for the tuna regional fisheries management organisations. *R package version 2.15, URL: http://cran.r-project.org/web/packages/kobe/kobe.pdf*.

Kingston, N. M., & McKinley, R. L. (1988, April). Assessing the structure of the GRE general test using confirmatory multidimensional item response theory. *Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA*.

Li, Y., & Rupp, A. A. (2011). Performance of the s-χ^2 statistic for full-information bifactor models. *Educational and Psychological Measurement*, *71*(6), 986–1005.

Martel, M. M., Von Eye, A., & Nigg, J. T. (2010). Revisiting the latent structure of ADHD: is there a “g”factor? *Journal of Child Psychology and Psychiatry*, *51*(8), 905–914.

Muthén, L., & Muthén, B. (1998-2012). Mplus. *The comprehensive modelling program for applied researchers: user’s guide*, *5*.

Navarro. (2013). lsr: Companion package to "learning statistics with r". *R package version 0.2, URL: http://cran.r-project.org/web/packages/lsr/lsr.pdf*.

Norris, M. (2010). *Examining the autism phenotype: The structure of autism spectrum disorders as measured by the autism diagnostic observation schedule*. Doctoral dissertation, retrieved from http://search.proquest.com.www2.lib.ku.edu:2048/docview/815352304?accountid=14556.

Patrick, C. J., Hicks, B. M., Nichol, P. E., & Krueger, R. F. (2007). A bifactor approach to modeling the structure of the psychopathy checklist-revised. *Journal of Personality Disorders*, *21*(2), 118–141.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reininghaus, U., McCabe, R., Burns, T., Croudace, T., & Priebe, S. (2011). Measuring patients' views: a bifactor model of distinct patient-reported outcomes in psychosis. *Psychological Medicine*, *41*(2), 277–289.

Reise, Moore, & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*, *71*(4), 684–711.

Reise, Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559.

Reise, S. P., Horan, W. P., & Blanchard, J. J. (2011). The challenges of fitting an item response theory model to the social anhedonia scale. *Journal of Personality Assessment*, *93*(3), 213–224.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31.

Reise, S. P., Ventura, J. S., Keefe, R., Baade, L. E., Gold, J. M., Green, M. F., … Seidman, L. J. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological Assessment*, *23*(1), 245–261.

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. *Educational Testing Service Research Report No: RR-09-37, ETS,* *Princeton, NJ*.

Rijmen, F. (2010). Formal relations and an empirical comparison among the Bi-Factor, the testlet, and a Second-Order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361–372.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61.

Seo, D. G. (2011). *Application of the bifactor model to computerized adaptive testing*. Doctoral dissertation, retrieved from http://search.proquest.com/docview/857080270?accountid=14556.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2007). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, *25*(7), E34–E46.

Statistics, I. I. S. (2010). Core system user's guide. *Chicago: SPSS Inc*.

Swineford, F. (1941). Some comparisons of the multiple-factor and the bi-factor methods of analysis. *Psychometrika*, *6*(6), 375–382.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the brief symptom inventory. *Psychological Assessment*, *24*(1), 101–113.

Thomas, M. L., & Locke, D. E. (2010). Psychometric properties of the MMPI-2-RF somatic complaints (RC1) scale. *Psychological Assessment*, *22*(3), 492–503.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*(2), 111–136.

Venables, W. N., Smith, D. M., & Team, R. D. C. (2002). *An introduction to r.* Network Theory.

Weekers, A. M. (2009). *Modeling typical performance measures* [Doctoral Dissertation]. retrieved from http://doc.utwente.nl/68828.

Wickham, H. (2008). ggplot2: An implementation of the grammar of graphics. *R package version 0.7, URL: http://CRAN. R-project. org/package= ggplot2*.

Yang, Y., Song, L., & Xu, T. (2002). Robust estimator for correlated observations based on bifactor equivalent weights. *Journal of Geodesy*, *76*(6), 353–358.

Appendix A

Theta Parameter Estimate Bias of the Validity Model

Table A.1: Theta Parameter Estimate Bias of the Validity Model

	$\overline{X}$	σ	Min	Q1	Q2	Q3	Max
P	0.020	0.237	-1.278	-0.126	0.024	0.159	1.859
S1	0.025	0.315	-1.995	-0.122	0.047	0.195	1.960
S2	0.008	0.382	-2.024	-0.200	0.000	0.223	2.046
S3	0.015	0.320	-2.225	-0.155	0.000	0.172	2.327
S4	0.014	0.411	-2.188	-0.225	-0.007	0.225	2.157

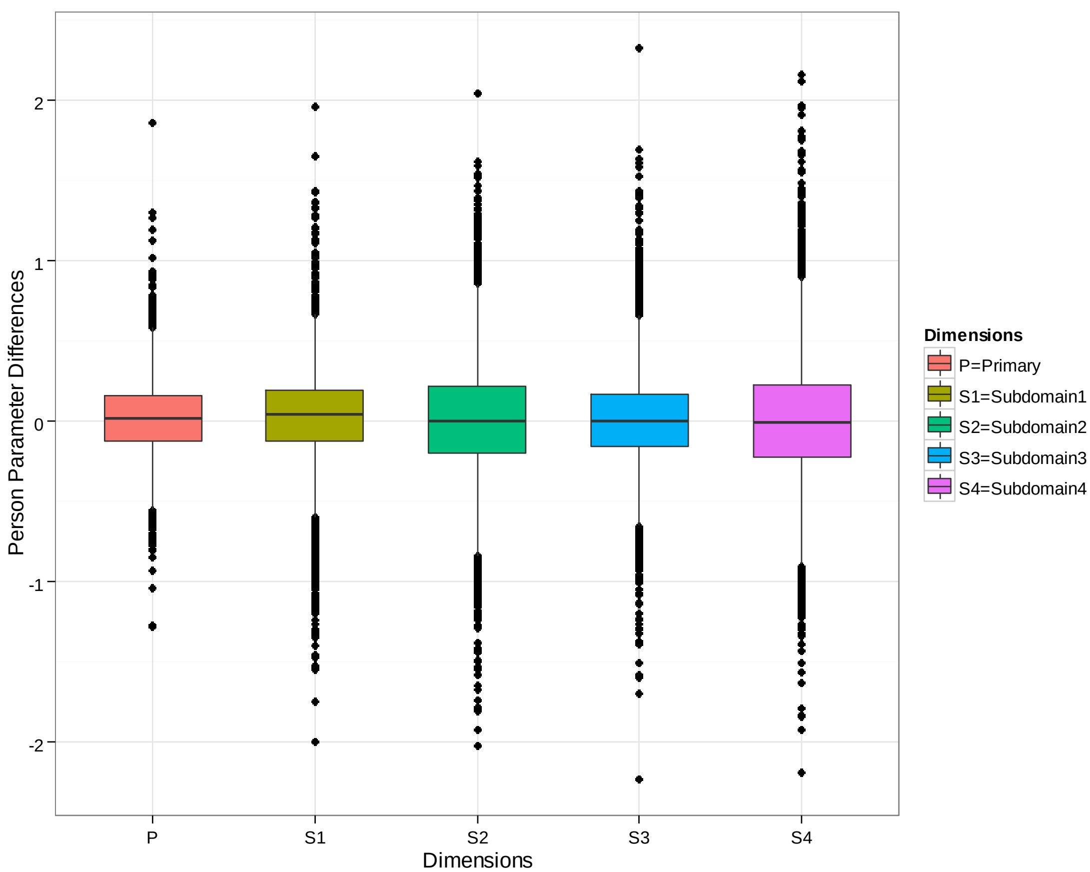

Figure A.1: Theta Parameter Estimate Bias of the Validity Model

Appendix B

True Discrimination and Intercept Parameters for the Simulated Data

Table B.1: True Item Parameters for 40 Items

	a_g	a_1	a_2	a_3	a_4	d
Item 1	1.476	0.854	0.000	0.000	0.000	2.198
Item 2	0.988	1.705	0.000	0.000	0.000	0.686
Item 3	0.560	0.702	0.000	0.000	0.000	0.469
Item 4	1.581	0.368	0.000	0.000	0.000	-2.066
Item 5	1.124	0.301	0.000	0.000	0.000	-2.124
Item 6	0.280	1.460	0.000	0.000	0.000	2.246
Item 7	1.460	1.918	0.000	0.000	0.000	-0.268
Item 8	1.363	1.025	0.000	0.000	0.000	1.298
Item 9	0.276	1.484	0.000	0.000	0.000	1.011
Item 10	0.394	1.647	0.000	0.000	0.000	-0.466
Item 11	0.903	0.000	1.445	0.000	0.000	1.743
Item 12	0.686	0.000	1.439	0.000	0.000	2.900
Item 13	1.354	0.000	1.692	0.000	0.000	1.448
Item 14	0.340	0.000	1.568	0.000	0.000	0.095
Item 15	0.699	0.000	0.288	0.000	0.000	-0.665
Item 16	1.417	0.000	0.730	0.000	0.000	-0.429
Item 17	1.704	0.000	0.325	0.000	0.000	0.035
Item 18	0.856	0.000	1.653	0.000	0.000	0.977
Item 19	0.333	0.000	0.761	0.000	0.000	1.170
Item 20	0.504	0.000	1.404	0.000	0.000	2.736
Item 21	1.309	0.000	0.000	1.175	0.000	0.278
Item 22	0.549	0.000	0.000	0.489	0.000	-0.554
Item 23	1.479	0.000	0.000	0.669	0.000	1.480
Item 24	1.717	0.000	0.000	0.514	0.000	-1.434

Item 25	0.756	0.000	0.000	0.287	0.000	-1.206
Item 26	1.370	0.000	0.000	0.948	0.000	1.826
Item 27	1.240	0.000	0.000	1.474	0.000	1.029
Item 28	0.663	0.000	0.000	1.326	0.000	2.107
Item 29	0.784	0.000	0.000	1.154	0.000	1.734
Item 30	0.862	0.000	0.000	0.934	0.000	-0.295
Item 31	0.845	0.000	0.000	0.000	1.871	3.541
Item 32	0.936	0.000	0.000	0.000	0.398	-0.626
Item 33	0.988	0.000	0.000	0.000	0.857	-0.951
Item 34	1.314	0.000	0.000	0.000	1.978	0.100
Item 35	0.600	0.000	0.000	0.000	1.231	-0.296
Item 36	1.329	0.000	0.000	0.000	0.705	-2.663
Item 37	0.419	0.000	0.000	0.000	0.404	-0.128
Item 38	1.463	0.000	0.000	0.000	1.465	-1.099
Item 39	0.872	0.000	0.000	0.000	1.935	-4.606
Item 40	1.026	0.000	0.000	0.000	0.930	4.063

Table B.2: True Item Parameters for 60 Items

	a_g	a_1	a_2	a_3	a_4	d
Item 1	1.340	1.214	0.000	0.000	0.000	0.383
Item 2	0.772	1.567	0.000	0.000	0.000	1.820
Item 3	0.634	0.624	0.000	0.000	0.000	1.026
Item 4	0.881	0.225	0.000	0.000	0.000	-0.293
Item 5	0.834	0.515	0.000	0.000	0.000	1.470
Item 6	0.736	0.562	0.000	0.000	0.000	0.413
Item 7	0.610	1.302	0.000	0.000	0.000	-2.493
Item 8	1.199	1.923	0.000	0.000	0.000	-1.158
Item 9	0.532	0.614	0.000	0.000	0.000	-0.081
Item 10	0.210	1.403	0.000	0.000	0.000	0.082
Item 11	0.864	1.399	0.000	0.000	0.000	2.866
Item 12	1.785	1.596	0.000	0.000	0.000	3.173
Item 13	0.754	1.994	0.000	0.000	0.000	1.168
Item 14	1.027	1.889	0.000	0.000	0.000	3.131
Item 15	0.975	0.364	0.000	0.000	0.000	-0.086
Item 16	0.939	0.000	0.583	0.000	0.000	-1.026
Item 17	1.937	0.000	0.503	0.000	0.000	1.435
Item 18	0.455	0.000	0.672	0.000	0.000	-0.781
Item 19	0.936	0.000	0.301	0.000	0.000	-1.520
Item 20	0.628	0.000	1.376	0.000	0.000	1.528
Item 21	0.432	0.000	1.364	0.000	0.000	-0.797
Item 22	0.776	0.000	1.618	0.000	0.000	-0.303
Item 23	1.254	0.000	1.727	0.000	0.000	-0.331
Item 24	1.392	0.000	1.192	0.000	0.000	-4.340

Item 25	1.804	0.000	1.802	0.000	0.000	4.044
Item 26	0.325	0.000	0.537	0.000	0.000	0.693
Item 27	1.903	0.000	1.735	0.000	0.000	-2.325
Item 28	1.547	0.000	1.551	0.000	0.000	-0.149
Item 29	1.381	0.000	1.220	0.000	0.000	2.117
Item 30	0.590	0.000	1.010	0.000	0.000	1.054
Item 31	0.906	0.000	0.000	1.235	0.000	1.835
Item 32	0.786	0.000	0.000	0.286	0.000	0.427
Item 33	1.074	0.000	0.000	0.527	0.000	-0.366
Item 34	1.726	0.000	0.000	1.965	0.000	0.178
Item 35	0.992	0.000	0.000	1.003	0.000	-0.525
Item 36	1.114	0.000	0.000	0.987	0.000	-1.703
Item 37	2.000	0.000	0.000	1.263	0.000	3.406
Item 38	0.747	0.000	0.000	1.840	0.000	1.479
Item 39	0.289	0.000	0.000	0.622	0.000	0.743
Item 40	0.801	0.000	0.000	0.342	0.000	0.794
Item 41	0.431	0.000	0.000	1.598	0.000	-0.420
Item 42	1.484	0.000	0.000	1.698	0.000	-0.124
Item 43	1.262	0.000	0.000	1.043	0.000	-1.423
Item 44	1.174	0.000	0.000	1.992	0.000	1.147
Item 45	0.969	0.000	0.000	1.924	0.000	-1.700
Item 46	1.237	0.000	0.000	0.000	1.588	0.097
Item 47	0.811	0.000	0.000	0.000	1.416	-0.463
Item 48	1.780	0.000	0.000	0.000	1.741	5.839
Item 49	0.796	0.000	0.000	0.000	1.812	-0.697
Item 50	0.971	0.000	0.000	0.000	0.348	0.760
Item 51	1.081	0.000	0.000	0.000	0.365	-0.175

Item 52	1.338	0.000	0.000	0.000	1.232	-0.849
Item 53	0.551	0.000	0.000	0.000	0.537	1.038
Item 54	1.159	0.000	0.000	0.000	0.641	0.491
Item 55	1.219	0.000	0.000	0.000	1.025	-3.262
Item 56	1.026	0.000	0.000	0.000	1.233	1.652
Item 57	1.874	0.000	0.000	0.000	1.035	1.732
Item 58	0.429	0.000	0.000	0.000	0.557	0.633
Item 59	1.239	0.000	0.000	0.000	0.265	-1.675
Item 60	0.684	0.000	0.000	0.000	0.827	2.042

Table B.3: True Item Parameters for 80 Items

	a_g	a_1	a_2	a_3	a_4	d
Item 1	1.430	1.839	0.000	0.000	0.000	-2.709
Item 2	1.771	1.785	0.000	0.000	0.000	1.473
Item 3	1.442	1.323	0.000	0.000	0.000	-3.493
Item 4	0.409	0.349	0.000	0.000	0.000	0.717
Item 5	0.551	1.963	0.000	0.000	0.000	0.911
Item 6	1.030	1.621	0.000	0.000	0.000	-1.095
Item 7	0.566	0.970	0.000	0.000	0.000	3.246
Item 8	1.264	1.791	0.000	0.000	0.000	1.905
Item 9	0.873	0.324	0.000	0.000	0.000	0.430
Item 10	0.454	1.646	0.000	0.000	0.000	0.949
Item 11	0.373	1.378	0.000	0.000	0.000	0.029
Item 12	1.465	1.358	0.000	0.000	0.000	0.300
Item 13	0.340	1.895	0.000	0.000	0.000	1.209
Item 14	0.623	0.630	0.000	0.000	0.000	-1.172
Item 15	1.927	1.816	0.000	0.000	0.000	4.027
Item 16	1.631	0.853	0.000	0.000	0.000	0.804
Item 17	0.896	0.545	0.000	0.000	0.000	-1.018
Item 18	1.957	0.552	0.000	0.000	0.000	-0.057
Item 19	0.372	1.354	0.000	0.000	0.000	0.121
Item 20	1.247	0.641	0.000	0.000	0.000	-0.545
Item 21	1.050	0.000	0.342	0.000	0.000	-0.262
Item 22	0.744	0.000	1.530	0.000	0.000	0.245
Item 23	0.399	0.000	1.860	0.000	0.000	-1.373
Item 24	1.229	0.000	0.438	0.000	0.000	-0.483

Item 25	1.623	0.000	0.663	0.000	0.000	0.424
Item 26	1.349	0.000	0.466	0.000	0.000	2.101
Item 27	0.760	0.000	0.285	0.000	0.000	0.484
Item 28	1.447	0.000	0.209	0.000	0.000	1.677
Item 29	1.414	0.000	0.545	0.000	0.000	3.751
Item 30	1.565	0.000	1.738	0.000	0.000	1.436
Item 31	1.714	0.000	1.639	0.000	0.000	0.512
Item 32	0.633	0.000	0.269	0.000	0.000	-1.094
Item 33	1.424	0.000	0.419	0.000	0.000	-2.310
Item 34	0.327	0.000	1.754	0.000	0.000	-1.977
Item 35	1.364	0.000	1.751	0.000	0.000	2.435
Item 36	0.384	0.000	1.320	0.000	0.000	2.558
Item 37	1.920	0.000	1.649	0.000	0.000	2.313
Item 38	1.547	0.000	0.431	0.000	0.000	-2.001
Item 39	0.470	0.000	1.024	0.000	0.000	-0.099
Item 40	0.754	0.000	1.178	0.000	0.000	-0.592
Item 41	1.834	0.000	0.000	1.337	0.000	1.857
Item 42	0.520	0.000	0.000	1.387	0.000	2.286
Item 43	0.697	0.000	0.000	0.916	0.000	-0.640
Item 44	0.929	0.000	0.000	1.102	0.000	0.532
Item 45	1.904	0.000	0.000	1.457	0.000	2.510
Item 46	0.556	0.000	0.000	1.573	0.000	-0.030
Item 47	0.439	0.000	0.000	1.428	0.000	-1.318
Item 48	1.115	0.000	0.000	1.644	0.000	-1.752
Item 49	0.330	0.000	0.000	1.604	0.000	-1.680
Item 50	1.235	0.000	0.000	1.364	0.000	0.701
Item 51	1.582	0.000	0.000	1.723	0.000	-2.571

Item 52	0.762	0.000	0.000	1.322	0.000	0.047
Item 53	0.297	0.000	0.000	1.542	0.000	-0.298
Item 54	1.350	0.000	0.000	0.805	0.000	-2.098
Item 55	1.334	0.000	0.000	0.868	0.000	-1.163
Item 56	1.682	0.000	0.000	0.379	0.000	-0.097
Item 57	1.598	0.000	0.000	0.755	0.000	-2.349
Item 58	1.192	0.000	0.000	1.546	0.000	0.796
Item 59	0.257	0.000	0.000	0.529	0.000	0.481
Item 60	0.652	0.000	0.000	1.271	0.000	-0.513
Item 61	1.614	0.000	0.000	0.000	0.307	-0.097
Item 62	0.698	0.000	0.000	0.000	1.598	0.120
Item 63	0.380	0.000	0.000	0.000	0.270	0.179
Item 64	1.132	0.000	0.000	0.000	0.791	1.207
Item 65	0.592	0.000	0.000	0.000	1.075	-1.468
Item 66	1.119	0.000	0.000	0.000	0.402	1.975
Item 67	1.865	0.000	0.000	0.000	1.374	-2.699
Item 68	0.723	0.000	0.000	0.000	0.530	0.959
Item 69	0.933	0.000	0.000	0.000	0.333	-0.900
Item 70	1.806	0.000	0.000	0.000	0.586	2.502
Item 71	1.514	0.000	0.000	0.000	1.820	-1.551
Item 72	1.630	0.000	0.000	0.000	0.673	1.164
Item 73	0.330	0.000	0.000	0.000	1.528	-1.551
Item 74	1.444	0.000	0.000	0.000	1.010	1.204
Item 75	1.273	0.000	0.000	0.000	0.504	-1.535
Item 76	1.618	0.000	0.000	0.000	1.710	-0.916
Item 77	0.876	0.000	0.000	0.000	1.248	0.230
Item 78	0.509	0.000	0.000	0.000	1.060	0.548

Item 79	1.649	0.000	0.000	0.000	1.112	2.707
Item 80	1.433	0.000	0.000	0.000	1.335	0.059

Appendix C

Discrimination and Intercept Parameter Estimates for the Real Data

Table C.1: Item Parameter Estimates of the Real Data

	a_g	a_1	a_2	a_3	a_4	d
Item 1	1.420	0.547	0.000	0.000	0.000	2.857
Item 2	1.274	0.684	0.000	0.000	0.000	2.418
Item 3	1.346	0.822	0.000	0.000	0.000	2.201
Item 4	1.491	0.615	0.000	0.000	0.000	2.053
Item 5	1.152	-0.064	0.000	0.000	0.000	1.706
Item 6	0.922	0.172	0.000	0.000	0.000	0.859
Item 7	1.609	-0.640	0.000	0.000	0.000	3.141
Item 8	1.297	0.694	0.000	0.000	0.000	1.555
Item 9	1.020	-0.016	0.000	0.000	0.000	1.405
Item 10	1.203	-0.227	0.000	0.000	0.000	1.789
Item 11	0.927	0.478	0.000	0.000	0.000	0.434
Item 12	1.092	0.420	0.000	0.000	0.000	0.953
Item 13	1.656	-1.396	0.000	0.000	0.000	2.056
Item 14	1.454	-0.134	0.000	0.000	0.000	0.777
Item 15	1.008	-1.146	0.000	0.000	0.000	0.535
Item 16	0.947	0.726	0.000	0.000	0.000	-0.181
Item 17	1.383	1.397	0.000	0.000	0.000	1.240
Item 18	1.106	0.008	0.000	0.000	0.000	-0.326
Item 19	1.001	-0.158	0.000	0.000	0.000	1.366
Item 20	1.170	0.590	0.000	0.000	0.000	1.514
Item 21	0.917	0.657	0.000	0.000	0.000	1.194
Item 22	1.063	0.541	0.000	0.000	0.000	0.124
Item 23	0.865	0.223	0.000	0.000	0.000	1.278
Item 24	0.838	-0.202	0.000	0.000	0.000	1.682

Item 25	0.926	-0.182	0.000	0.000	0.000	1.047
Item 26	0.985	0.836	0.000	0.000	0.000	0.700
Item 27	0.346	-0.070	0.000	0.000	0.000	0.130
Item 28	1.093	0.138	0.000	0.000	0.000	0.703
Item 29	1.384	0.000	0.365	0.000	0.000	2.383
Item 30	1.488	0.000	0.670	0.000	0.000	2.344
Item 31	1.354	0.000	0.629	0.000	0.000	2.342
Item 32	1.169	0.000	-0.058	0.000	0.000	1.258
Item 33	1.282	0.000	0.212	0.000	0.000	2.022
Item 34	1.245	0.000	0.413	0.000	0.000	0.902
Item 35	0.684	0.000	0.708	0.000	0.000	2.078
Item 36	1.412	0.000	-0.059	0.000	0.000	2.615
Item 37	0.864	0.000	-0.145	0.000	0.000	0.164
Item 38	1.430	0.000	-0.128	0.000	0.000	0.833
Item 39	1.474	0.000	2.650	0.000	0.000	2.355
Item 40	0.998	0.000	1.417	0.000	0.000	0.901
Item 41	1.000	0.000	-1.143	0.000	0.000	0.792
Item 42	2.047	0.000	-1.928	0.000	0.000	2.493
Item 43	1.255	0.000	-1.710	0.000	0.000	0.545
Item 44	1.207	0.000	-0.389	0.000	0.000	0.370
Item 45	2.412	0.000	-2.955	0.000	0.000	3.292
Item 46	1.395	0.000	0.000	0.525	0.000	1.734
Item 47	1.865	0.000	0.000	1.308	0.000	2.942
Item 48	1.076	0.000	0.000	0.393	0.000	1.093
Item 49	2.011	0.000	0.000	0.993	0.000	2.702
Item 50	1.623	0.000	0.000	0.943	0.000	1.404
Item 51	1.399	0.000	0.000	-0.367	0.000	2.873

Item 52	1.053	0.000	0.000	0.349	0.000	1.466
Item 53	1.336	0.000	0.000	-0.554	0.000	2.522
Item 54	1.235	0.000	0.000	0.516	0.000	1.175
Item 55	1.678	0.000	0.000	-0.559	0.000	1.644
Item 56	1.201	0.000	0.000	-0.712	0.000	1.672
Item 57	1.051	0.000	0.000	0.152	0.000	1.571
Item 58	1.050	0.000	0.000	0.017	0.000	0.251
Item 59	1.024	0.000	0.000	0.358	0.000	1.763
Item 60	1.144	0.000	0.000	0.276	0.000	1.387
Item 61	0.862	0.000	0.000	0.296	0.000	0.975
Item 62	1.096	0.000	0.000	-0.203	0.000	1.355
Item 63	0.875	0.000	0.000	-0.063	0.000	0.675
Item 64	1.251	0.000	0.000	0.000	0.081	1.716
Item 65	1.125	0.000	0.000	0.000	-0.554	1.351
Item 66	1.717	0.000	0.000	0.000	-2.359	1.746
Item 67	1.127	0.000	0.000	0.000	0.252	0.796
Item 68	1.389	0.000	0.000	0.000	0.152	1.198
Item 69	1.102	0.000	0.000	0.000	0.443	0.524
Item 70	1.124	0.000	0.000	0.000	-0.275	1.249
Item 71	1.234	0.000	0.000	0.000	-0.548	2.544
Item 72	1.646	0.000	0.000	0.000	-0.948	2.710
Item 73	1.356	0.000	0.000	0.000	0.092	2.226
Item 74	1.470	0.000	0.000	0.000	0.134	1.564
Item 75	1.247	0.000	0.000	0.000	0.691	1.523
Item 76	1.149	0.000	0.000	0.000	2.169	0.147
Item 77	1.314	0.000	0.000	0.000	-0.263	0.923
Item 78	0.904	0.000	0.000	0.000	-0.199	1.853

Item 79	1.095	0.000	0.000	0.000	-0.068	1.418
Item 80	1.458	0.000	0.000	0.000	-0.791	1.025
Item 81	0.316	0.000	0.000	0.000	-1.669	1.107
Item 82	1.118	0.000	0.000	0.000	-0.019	1.047
Item 83	0.867	0.000	0.000	0.000	0.744	1.302
Item 84	0.648	0.000	0.000	0.000	-1.331	-0.052
Item 85	0.796	0.000	0.000	0.000	0.258	0.982
Item 86	0.875	0.000	0.000	0.000	0.425	0.509

Note. a_g = discriminations for math; a_1 = discriminations for number; a_2 = discriminations for data; a_3 = discriminations for geometry; a_4 = discriminations for algebra

Appendix D

R Code for Data Generation and Mplus Code for Factor Correlations of the Bifactor Model

The following was the example *R* code to generate the bifactor data of 40 items.

```
library(mvtnorm)
N <- 5000
Cmax <- 0.2
D <- 5
npb <- 10
M <- (D-1)*npb
meanTheta <- rep(0, D)
sdTheta<- rep(1, D)
rhoTheta<-matrix( c(1,rep(0,4),0,1,rep(0,3),0,0,1,rep(0,2),0,0,0,1,rep(0.4,1),rep(0,3),0.4,1),ncol=5)
covTheta<- rhoTheta * sdTheta %o% sdTheta
theta<- round(rmvnorm(N, mean = meanTheta, sigma = covTheta, method= "chol"),digits=3)
write.csv(theta, file="theta40.csv", row.names=FALSE)
set.seed(30)
diffp <- round(rnorm(M, 0, 1),digits=3)
set.seed(50)
discraw <- round(matrix(runif(D*M, min=0.2, max=2.0), ncol=D),digits=3)
blotterMatrix <- matrix(0, nrow=M, ncol=D-1) for ( i in 1:(D-1)){ blotterMatrix[(1 + (i-1)*npb):
(i*npb), i] <- 1 }
blotterMatrix <- cbind(1, blotterMatrix)
discp <- discraw * blotterMatrix
discssum<-sqrt(matrix(apply(discp*discp,1,sum),ncol=1))
DiffAndIntercept<-cbind(discp,intercept)
write.csv(DiffAndIntercept, file="DI40.csv", row.names=FALSE)
for(i in 1:200){items<-t(apply(invlink,2,function(col) rbinom(n=M, pr=col, size=1)))
myname<-paste("two-parameter", i, ".txt",sep="")
write.table(items, file=myname,row.name=FALSE, col.names=FALSE)}
```

The following was the Mplus code to study the factor correlations of the bifactor model.

```
TITLE: Target Oblique Rotation bifactor
DATA: FILE IS "*.txt";
VARIABLE: NAMES ARE R1-R86;
USEVARIABLE ARE R1-R86;
CATEGORICAL ARE R1-R86;
ANALYSIS:
ESTIMATOR = wls;
Rotation=target;
MODEL:
F BY R1-R86 (*1);
F1 BY R1-R28 R29-R86~0 (*1);
F2 BY R29-R45 R1-R28~0 R46-R86~0 (*1);
F3 BY R46-R63 R1-R45~0 R64-R86~0 (*1);
F4 BY R64-R86 R1-R63~0 (*1);
```